POR UNA APROXIMACIÓN HUMANISTA NO REACCIONARIA A LA IA

Towards a Non-Reactionary Humanistic Approach to AI

Javier Jurado González
Universidad Pontificia Comillas
jjuradog@icai.comillas.edu; https://orcid.org/0000-0002-4241-3466

Recibido: 27 febrero 2024
Aceptado: 4 marzo 2024
DOI: https://doi.org/10.14422/ryf.vol287.i1463.y2023.001

RESUMEN: Muchas publicaciones están tratando de pinchar la burbuja de expectativas sobre el desarrollo reciente de la IA. Sin embargo, en ciertos círculos humanistas donde un escepticismo conservador con respecto a las novedades va de serie, es conveniente detener la reflexión en el exceso contrario: la complacencia de posturas que se contentan con argumentos débiles y conceptos periclitados ya superados en la literatura científica y filosófica. Estas perspectivas tienden a adoptar una postura reaccionaria ante la IA, aferrándose defensivamente a cualquier argumento que pueda preservar la singularidad humana a costa de renunciar a una cierta honestidad intelectual. Así afirma taxativamente que la IA nunca logrará alcanzarla. No obstante, es posible adoptar posturas humanistas receptivas a los desarrollos de la IA, abiertas a sus retos actuales y capaces de dialogar y refinar sus argumentos a través de algunas claves: dignificar la miserabilidad humana, tender puentes interdisciplinares, y mantener la prudencia, la cortesía y la suspensión del juicio cuando sea preciso.
PALABRAS CLAVE*:* filosofía de la inteligencia artificial, filosofía de la mente, humanismo tecnológico, fundamentación semántica, problema difícil de la consciencia.

ABSTRACT: Numerous publications are engaged in deflating the inflated expectations surrounding recent advancement in AI. However, within certain humanist circles where conservative skepticism regarding novelties is prevalent, it is convenient to reflect on the converse extreme: the complacency of viewpoints that rely on weak arguments and outdated concepts which have been surpassed in both scientific and philosophical literature. These perspectives tend to adopt a reactionary stance towards AI, defensively holding onto any assertion that preserves human uniqueness, even at the expense of intellectual integrity. Such a defensive posture steadfastly asserts that AI will never attain human singularity. Nevertheless, it is conceivable to embrace humanistic perspectives that remain receptive to AI's advancements, addressing its current challenges while engaging in dialogue and refining arguments. This can be achieved through several approaches, including digni-

fying human misery, fostering interdisciplinary collaboration, and exercising prudence, courtesy, and the suspension of judgment when warranted.
KEYWORDS: philosophy of artificial intelligence, philosophy of mind, technological humanism, semantic grounding, hard problem of consciousness.

1. **LA ACTITUD REACCIONARIA**

Decía Bonhoeffer que no podemos centrarnos en buscar a Dios en lo que no conocemos y que la ciencia no ha sido aún capaz de explicar, utilizándolo para tapar los agujeros de nuestra ignorancia. Porque eso supone ir arrinconándolo en un estatus ontológico un tanto sospechoso, en una posición tan huidiza que pudiera parecerse a la de la inexistencia. En su lugar, abogaba por abandonar la pereza intelectual, remozar y hacer más sutiles nuestros argumentos y encontrar a Dios *en lo que sí conocemos* (Bonhoeffer, 2001, p. 218).

Del mismo modo, parece existir un cierto acorralamiento de lo que significa ser humano ante los nuevos desarrollos de la Inteligencia Artificial (IA). Y entre las múltiples respuestas, existe un cierto tipo reaccionario, una forma conservadora de entender el humanismo cristiano, que responde defensivamente tratando de proteger al hombre de dos formas: la primera es subrayar ante todo lo mucho que desconocemos, los límites y las carencias del desarrollo de la ciencia, la tecnología o la propia filosofía, tratando de preservar la esencia humana *quia absurdum* en un reducto de lo desconocido. La segunda forma entiende la exhortación de Bonhoeffer y procura hallar, en lo que sí conoce, fundamentos para asegurar la singularidad humana. Sin embargo, la urgencia ideológica precipita respuestas atrevidas y con frecuencia alumbra argumentos inválidos que echan mano de conceptos e ideas que el tiempo ha ido deslavazando. Esto se da especialmente en algunos ámbitos, en los que se desconocen en gran medida los fundamentos y las reflexiones que desde hace décadas se han vertido sobre la filosofía de la IA, la filosofía de la mente, las investigaciones científicas en torno a las neurociencias, la psicología o la biología evolutiva y los desarrollos conceptuales en torno a la IA[1]. Este desconocimiento acaba encastillándose y mostrando

[1] Entre muchos otros, convendría echar un vistazo a las principales obras al respecto de autores como Alan Turing (1912), Herbert A. Simon (1916), Walter Pitts (1923), Hilary Putnam (1926), John McCarthy (1927), Marvin Minsky (1927), Hubert Dreyfus (1929), Roger Penrose (1931), John Searle (1932), Jerry Fodor (1935), Judea Pearl (1936), Thomas Nagel

un narcisismo herido que resuena a otros episodios, porque la historia no se repite, pero rima, como le atribuyen haber dicho a Twain. Aunque Freud se tuviera excesiva autoestima, cabe reconocerle algún papel en esa tendencia que describió hablando de las *tres humillaciones del narcisismo humano*: desbancado del centro del Universo por Copérnico, destronado de la cumbre de la naturaleza por Darwin, y desprovisto del control de su consciencia por el famoso psicoanalista (Freud, S., 1968, vol. II, pp. 1110-1112). Como si de una nueva ola se tratase, parece que la IA no sólo está llamada a pulverizar nuestras marcas y derrotarnos en juegos milenarios como el ajedrez o el go, sino que empieza a cuestionar qué es lo que somos realmente, qué es lo que como humanos nos seguirá distinguiendo de ella y evitará que nunca podamos ser reemplazados por su ascenso. Entonces, entre otras, se produce esta respuesta un tanto reaccionaria que, intoxicada a priori de cierta animadversión, se pregunta prejuiciosamente ¿de qué hablamos realmente cuando decimos *inteligencia*?

Sin duda, un escepticismo sano sobre las grandilocuentes declaraciones en torno a la IA resulta pertinente. Así se han pronunciado diversos autores de renombre que alertan sobre los peligros de la *mitificación de la IA*. Esta mitología no sólo capta financiación y atención de forma tramposa y sensacionalista (e incluso catastrofista), sino que también podría llevarnos a una complacencia que se empeñe en persistir en el desarrollo de las líneas tecnológicas estrechas contemporáneas e ignore el *misterio científico* que aún sigue siendo la inteligencia humana, sin enfrentarlo directamente, desanimando a los científicos a pensar en nuevas formas de abordar el reto de la inteligencia.

Por ejemplo, investigaciones recientes contrastan los mecanismos de aprendizaje del cerebro humano con los del *deep learning* de la IA: el cerebro realiza tareas de clasificación complejas con tanta eficacia como la IA con muchas menos capas, siguiendo una dinámica más lenta y ruidosa pero energéticamente mucho más eficiente e indican la necesidad de un cambio tecnológico si se pretende imitar mejor la estructura del cerebro y sus métodos de aprendizaje (Tevet *et al.*, 2024). Una cultura saludable de la investigación pondría foco en las incógnitas y no en exagerar las capacidades de los métodos existentes (Larson, 2021). Las grandes empresas tecnológicas

(1937), Daniel Dennett (1942), Patricia Churchland (1943), Douglas Hofstadter (1945), Ray Kurzweil (1948), Ronald J. Brachman (1949), Andy Clark (1957), Yann Le Cun (1960), Jaron Lanier (1960), Stuart Russell (1962), Yoshua Bengio (1964), Luciano Floridi (1964), D. Chalmers (1966), Nick Bostrom (1973), etc.

están captando el talento en una reñida carrera por sacarle partido económico a la IA actual, hurtando a la investigación la exploración de otros enfoques alternativos a más largo plazo, lo que podría estancar los progresos hasta la fecha y llevarnos a un nuevo invierno, encerrados en modelos que podrían no tener salida.

Sin embargo, floreciendo anejas a estas reservas razonables, surgen otro tipo de respuestas que sobrerreaccionan ante la potencial amenaza de este artefacto humano. No sólo al nivel laboral, como luditas, o al nivel pragmático-existencial para los más *singularistas*[2]. Sino también al nivel más metafísico del reconocimiento de la dignidad humana, porque la IA comienza a ofrecer comportamientos que cuestionan nuestra unicidad, nuestra exclusividad, y en cierta forma hieren nuestro orgullo. No es casualidad que algunas de estas posturas conservadoras coqueteen con alarmismos que conectan con la visión distópico-apocalíptica sobre la IA que predomina en la ciencia ficción y en el sensacionalismo mediático actual (Oviedo, 2022). Pero otras de ellas también tienden a infravalorar el impacto que la IA va a producir próximamente.

El humanismo ha vivido a lo largo de la historia importantes convulsiones, pero con el tiempo ha sido capaz de ir asimilando con madurez no ser el centro del universo ni de la naturaleza, y estar constreñido por las circunstancias materiales y el subconsciente. Y a pesar de todo, afirmarse, como un valor. Sin embargo, para resistir así, al menos retenía esa diferencia específica de ser el único *Sapiens*. No obstante, los más *tecnooptimistas*, aunque admitan que es un producto de su mano, prometen que la IA acabará rompiendo esa exclusividad. Y los resultados prácticos comienzan a avanzar en esa dirección despertando las alarmas de muchos, que se ponen a la defensiva. Esto sucede particularmente en ciertos círculos humanistas conservadores, donde la innovación tecnológica suele suscitar temores y críticas porque se fundamentan en gran medida en la tradición (Oviedo, 2022). Indudablemente, las teorías científicas de Copérnico, Darwin o Freud han recibido matizaciones o correcciones, pero el impacto emocional que provocaron en muchos humanistas fue innegable. Ello les exigió ese crecimiento, esa madurez. Aunque la IA esté todavía muy lejos en diversos sentidos de alcanzar nuestras capacidades y quizá nunca las alcance, los humanistas, ¿no deberíamos haber aprendido de aquellas experiencias para

[2] En alusión a los más preocupados por el advenimiento inminente de la *singularidad* pronosticada, entre otros, por R. Kurzweil, como en Kurzweil (2005).

seguir poniendo en valor al hombre sin reaccionar torpemente ante los nuevos avances?

Para acoger y saber ponderar el desarrollo de la IA —ya sea que nos lleve a un nuevo invierno o prosiga en su sorprendente revolución actual—, una reflexión humanista seria no puede consentirse recurrir a argumentos o conceptos periclitados, filosóficamente superados o al menos profundamente desacreditados, desconociendo los debates que se han revisitado reiteradamente en la literatura. Parece recomendable abandonar ese *humanismo fácil* que opone al hombre y a la máquina, recubriéndose de ignorancia o resentimiento (Simondon, 2007, p. 31). La reacción ante los nuevos desarrollos no puede ser una retirada al monte, tratando sistemáticamente de resistir un supuesto asedio, como los famosos galos en su rincón bretón ante Roma, obstinándose en encontrar qué es aquello que radicalmente nos distingue como seres extraordinarios en la naturaleza, y cuáles son las muchas carencias que la nueva tecnología tiene y presuntamente siempre tendrá. Al menos, no es prudente hacerlo mediante afirmaciones taxativas que el tiempo acostumbra a derrumbar, y menos ante dos realidades —una antigua, el hombre; y otra nueva, la IA— que en gran medida desconocemos en profundidad.

Se quejaba Voltaire de la actitud de ese teólogo que, incapaz de reconocer el poder de la razón, o de la ciencia hoy acusada con excesiva frecuencia de "cientifista" en estos círculos, arremete contra ella señalando sus limitaciones, en aquella iluminadora frase: "*Sólo tenemos una luz, la razón. Viene el teólogo, dice que alumbra poco, y la apaga*" (Cit. por Fraijó, 1998, pp. 22-46). No es posible hoy hacer reflexión filosófica ni teológica por más humanista que se precie con honestidad intelectual sin escuchar lo que las ciencias tentativa y coherentemente nos están aportando sobre nuestra realidad humana y tecnológica, y hacerlo desde una actitud crítica pero abierta y humilde. Porque, si bien no todos los teólogos ni todos los filósofos apagan la luz, hay unos cuantos que se pronuncian sobre las más diversas disciplinas, como si la especialización, a la que los niveles de conocimiento actuales nos obligan, no fuera con ellos. Al hacerlo, en seguida revelan una escasa comprensión del asunto del que hablan, ya sea de los fundamentos de nuestra biología evolutiva, de la psicología cognitiva, de las neurociencias, o de los últimos logros en tecnología de IA. La filosofía, ciertamente, debe asomarse a todo —más que "comprenderlo" todo, como quisiera Lachelier—, pero siempre presidida por el mandato socrático "sólo sé que no sé nada", enfocándose ante todo en las preguntas pertinentes. Otra cosa es

incurrir en aquello que José Gaos llamaba la *soberbia luciferina* del filósofo (Gaos, 1982, pp. 119-120).

Para comprobar si existen o si podrán existir máquinas inteligentes, hoy podemos reconocer que el *imitation game* conocido como el test de Turing (Turing, 1950) parece una prueba insuficiente, aunque necesaria, porque se limita a evaluar la conducta externa, y el conductismo en el pasado ya nos ha ofrecido importantísimas limitaciones para una comprensión exhaustiva, particularmente de lo que significa ser *inteligente*. Volveremos sobre ello. Pero hagamos antes caso del propio Turing, y desprendámonos del prejuicio que supone ir buscando a toda costa un cierto tipo de *consuelo* (*comfort*) que refrende la excepcionalidad humana: *"It is customary* [...] *to offer a grain of comfort, in the form of a statement that some peculiarly human characteristic could never be imitated by a machine.* […] *I cannot offer any such comfort"* (Turing, 1951, p. 486).

2. **ACUDIR AL FRENTE DESARMADO**

No puede negarse que la concepción que tenemos de nosotros mismos sufre hoy una serie de desafíos constantes que son tremendamente interesantes para la investigación. Los argumentos que tratan de delimitarlos —en el sentido griego de concretar y aclarar— levantan. Pero las razones que en algunos se esgrimen resultan un tanto desoladoras, porque apenas resistirían el primer embate dialéctico. Este tipo de argumentos pretenden cubrir la desnudez del emperador, pero en realidad lo dejan desarmado. Para advertirlos, esta sección se toma la licencia de señalar algunos de ellos sin identificar por cortesía a sus autores, como muestra genérica que invite a la reflexión sobre el fondo.

2.1. La impugnación etimológica

Uno de estos argumentos pobres es el de la *impugnación etimológica*. Una forma de tratar de negar que la IA pueda algún día alcanzar un nivel de inteligencia humana es acudir a los supuestos arcanos profundos del significado originario de la voz *inteligencia* que, oh casualidad, siempre se encontrarán íntima e inseparablemente unidos a una antediluviana y misteriosa naturaleza o esencia humana. Como si los romanos del *inte-lligere* retuvieran la privilegiada llave secreta para determinar si nuestras redes neuronales

convolucionales o nuestros *transformers* podrán avanzar hacia el desarrollo de comportamientos que pudiéramos considerar inteligentes.

Ciertamente, puede que tenga algún sentido acudir a la raíz etimológica y con ella a nuestra historia para ayudarnos a suscitar perspectivas, a rescatar reflexiones, dimensiones o conceptualizaciones de nuestra inteligencia que enriquezcan las dimensiones del espacio multidimensional con el que tratamos de definirla. Por eso, cabe rechazar la suficiencia e imprudencia de un desmedido *esprit de géométrie* pascaliano, heredado en la fría racionalidad tecnocientífica e instrumental. Hay mucho del *esprit de finesse* que probablemente siempre se muestre renuente a dejarse atrapar por nuestro esfuerzo científico. La riqueza de la inteligencia humana sigue siendo un arcano poliédrico, y está repleta de dimensiones y matices. La teoría de las inteligencias múltiples ha sido severamente cuestionada por su falta de evidencia científica y a pesar de todo persiste en el imaginario colectivo (Waterhouse, 2023), probablemente porque la plasticidad de nuestra inteligencia muestra distintas habilidades y se adapta a múltiples facetas que nos permiten hablar en sentido figurado de "inteligencia cordial", "inteligencia emocional", e incluso de "inteligencia espiritual" de la que probablemente habrían hablado nuestros místicos. La IA tiene un reto enormemente grande por delante para poder aproximarse a esa riqueza. Volveremos sobre esto.

Pero creer, en pleno siglo XXI, que las esencias sustanciales más profundas se encuentran ocultas en nuestros lexemas antediluvianos resulta algo tan deslumbrante como infundado. Es necesario escuchar a las neurociencias y a la psicología, a la filosofía de la mente contemporánea, e indagar en y con ellas para preguntarse qué sabemos de la inteligencia, de dónde viene, cómo surgió, cómo funciona. Sin perder de vista que, aproximándonos a un ser vivo como el ser humano, es difícil que algo tenga sentido en él si no es a la luz de la evolución (Dobzhansky, 2013). Y que, por más que creamos haber encontrado algo genuinamente inexplicable que parezca escapar a la explicación científica, la evolución demuestra una y otra vez, como decía Orgel en su segunda ley, que suele ser más lista que lo que nuestra imaginación da de sí (Sejnowski, 2018, p. 246), y que por tanto las explicaciones naturalistas acaban sistemáticamente ganando la partida.

2.2. Emociones, aprendizaje y enseñanza

Resultan un tanto ingenuos los argumentos que siguen afirmando una suerte de excepcionalidad antrópica que coquetea en exceso con el discurso más

creacionista, científica y reiteradamente desacreditado. O una excepcionalidad naturalista que ignora el gradiente continuo que existe entre las especies naturales y la nuestra, y que la IA podría de alguna forma explorar y, llegado el caso, transitar.

Indudablemente, la naturaleza ha perfilado durante millones de años la inteligencia humana dentro de una psicología tan enormemente rica que en algunas de sus competencias la IA se encuentra todavía a una distancia abismal. Del mismo modo que, probablemente, nunca podremos llegar a experimentar lo que significa ser un murciélago al no contar con el conexionado neuronal exacto ni la experimentación directa de su sistema sensorial (Nagel, 1980), es posible que la IA, especialmente tal y como hoy se construye, tenga vedado un camino en algunos de los procesos cognitivos básicos como para desarrollarlos plenamente y que podamos llegar a considerarla como realmente inteligente. Últimamente, parecemos verla despuntar en la percepción, en el aprendizaje, en la atención, la memoria o el lenguaje, pero quizá, como luego comentaremos, tenga limitaciones insoslayables en el desarrollo de la motivación, del pensamiento o de la emoción.

No obstante, en lugar de acudir a los increíbles desafíos que supone el problema de la fundamentación semántica o el problema difícil de la consciencia y de los *qualia*, que luego comentaremos, se afirma sin sonrojo que sólo los seres humanos somos capaces de desarrollar algunos de estos procesos cognitivos, como por ejemplo *aprender por observación*, o dedicar tiempo activo a *enseñar*. O se acude al bastión de las *emociones*, como una fortaleza inexpugnable del fenómeno humano al que jamás accederán las máquinas. Como si sobre esta fortaleza pudiera parafrasearse el frontispicio el de la Academia platónica: "*No entre el que no sea humano*". Sorprende que a estas alturas no se haya asumido que en la misma naturaleza existen mecanismos de altruismo de parentesco y de altruismo recíproco, articulados mediante las emociones y operados por los genes, que hacen adaptativos los comportamientos de cooperación y aprendizaje que muestran diversas especies, particularmente las especies sociales. No, no somos los únicos que tenemos emociones, ni tampoco los únicos que aprendemos por observación o que dedicamos tiempo a enseñar.

Hoy es prácticamente imposible discutir que las emociones son adaptaciones biológicas (Nesse y Ellsworth, 2009). Ya Darwin argumentó que existe continuidad entre la vida emocional de los humanos y la de otros animales, y la evidencia a su favor es cada vez mayor sin que a la vez pueda apreciarse una frontera clara entre los animales que las sentirían y los que no (Bekoff, 2000). En primates, esto se refleja incluso en su activación cerebral, semejante a la

humana, con quien ancestros comunes que fueron los que con toda probabilidad comenzaron a desarrollarlas (Lindell, 2013). Los desarrollos en IA enfocan las emociones, hoy en día, eminentemente desde el punto de vista que pretende mejorar su capacidad para interpretarlas en la interacción humano-máquina (Lee *et al.*, 2024). No cabría pensar en desarrollar máquinas a las que quisiéramos llamar inteligentes que no fueran capaces de interpretar correctamente nuestras emociones. Pero en lugar de adentrarnos en la intrincada cuestión de si es realmente posible llegar a interpretarlas correctamente sin *sentir* esas mismas emociones, y redirigir el problema al de la consciencia y la experiencia subjetiva, se recurre a la simple afirmación de que, salvo los excepcionales seres humanos, nada más en este universo podrá tener nunca ira, tristeza o alegría.

Desde el punto de vista del aprendizaje por observación y de la enseñanza activa, una postura científicamente honesta no se permitiría aseverar que el ser humano es cualitativamente único. Así, por ejemplo, las hormigas, las abejas, los estorninos, los cuervos, los lobos, los elefantes, o los chimpancés tienen un comportamiento instintivo relativamente bajo y, aunque su capacidad cultural con mucha frecuencia se restrinja al grupo de pertenencia, sin ella los individuos resultan inviables, tal y como han mostrado los múltiples intentos por reintroducir en sus hábitats naturales muchos ejemplares criados en cautividad (Brakes *et al.*, 2019). Muchos animales pueden aprender comportamientos por imitación, como los abejorros (Bridges *et al.*, 2024). Particularmente lo hacen algunos mamíferos, como por ejemplo los chimpancés (Van Leeuwen et al., 2024), que son capaces de comunicarse sofisticadamente (Gabrić, 2021) y de desarrollar una cultura imprescindible para su supervivencia (Mosterín, 1998, pp. 146-152). De hecho, los chimpancés llegan con ella a *automedicarse* y a aplicar *técnicas curativas* a sus congéneres (Mascaro et al., 2022). Pero esta capacidad no es exclusiva ni siquiera de los primates como nosotros: las hormigas han dado muestras de *enseñanza interactiva*: las experimentadas con el rol de *instructoras* conducen a sus compañeras inexpertas *aprendices* hasta la fuente de alimento, adecuando su velocidad a la del aprendizaje de la anterior (Franks & Richardson, 2006).

Sabiendo esto del mundo natural, ¿no deberíamos celebrar como un logro la capacidad que hemos tenido para trasladar ciertos mecanismos de aprendizaje biológico a las redes neuronales y no ningunearlos como si ostentáramos su monopolio? Lo más honesto es admitir que resultan abrumadoras las capacidades de aprendizaje que, al estilo biológico, han desarrollado los mecanismos de IA acudiendo no sólo a la supervisión humana sino también al entrenamiento por refuerzo o el autoaprendizaje. Y observar con interés

las propuestas que dicen que los sistemas de IA podrían mejorarse mediante el uso de modelos de autoaprendizaje biológicamente más plausibles inspirados en la neurociencia y el cerebro (Bengio *et al.*, 2016; Hassabis *et al.*, 2017; Ullman, 2019).

Los defensores categóricos de la exclusividad de la inteligencia humana no son capaces de admitir el gradiente palpable que la ciencia nos está proporcionando con sus descubrimientos. La inteligencia humana no representa una excepción aislada, sino que revela una diferencia de grado en la naturaleza, pero no sustancial o cualitativa frente a la que exhiben otras especies, especialmente las que tienen un alto cociente de encefalización. Por ejemplo, los delfines son capaces del aprendizaje vocal, del etiquetado referencial, de la comprensión de cierta sintaxis, de la atención conjunta, de un diverso lenguaje gestual, y de un altísimo nivel de reconocimiento individual, incluso con identificadores únicos, al estilo de los nombres propios (Janik, 2013). Los pulpos o algunas cacatúas, a pesar de no haber evolucionado utilizando herramientas, cuando son expuestas a potenciales utensilios presentan la capacidad de adopción, de innovación y de planificación técnica, empleando herramientas múltiples para resolver problemas complejos (Osuna-Mascaró, A. J., 2022). Como estos, existen otros múltiples ejemplos en cuervos, buitres, urracas, elefantes y por supuesto primates como los chimpancés o los bonobos.

La naturaleza, por tanto, muestra que pueden desarrollarse inteligencias más limitadas a partir de cerebros con otras características que los hacen probablemente menos complejos o que cuentan con una densidad menor de neuronas en su corteza, pero que evolutivamente han explorado el crecimiento de las capacidades inteligentes como forma adaptativa. En esa línea, cabe recordar que existieron en nuestra rama evolutiva, como miembros legítimos de nuestro género, otras especies humanas que finalmente se extinguieron y que exhibieron comportamientos aún más inteligentes, incluyendo el dominio de industrias líticas o del fuego hace más de 400.000 años (MacDonald *et al.*, 2021), el dominio de una teoría de la mente, de la imitación e incluso del lenguaje como en el caso de los Neandertales (Roth y Dicke, 2005), los cuales es posible que llegasen a adoptar comportamientos simbólicos similares a los de los humanos modernos (Aubert *et al.*, 2018), pudiendo ser autores de las primeras expresiones artísticas (Appenzeller, 2018).

Observando este muestrario, no parece que tenga sentido rechazar dogmáticamente la posibilidad de que un día la evolución pudiera acabar alumbrando otra especie que pudiéramos calificar de inteligente, usurpándonos el pretendido trono. Ciertamente, la IA se encuentra desarrollando unas capacidades

que, por más que se basen en redes de nodos llamadas "neuronales", en su substrato han sido hasta ahora desarrolladas de una forma enormemente diferente a las de los cerebros vivos. Efectivamente, las funciones autoorientadas o los mecanismos evolutivos en términos estrictamente biológicos no son directamente aplicables a los sistemas tecnológicos. De hecho, el mecanismo biológico parece radicalmente diferente de cualquier máquina análoga que hasta ahora se haya propuesto (Herzog & Herzog, 2024). Sin embargo, trasladando la cuestión, ¿por qué oponerse de antemano y taxativamente a que el desarrollo de la IA pueda seguir progresando en sus capacidades para alcanzar la suficiente complejidad estructural y energética, contando con la suficiente capacidad de cómputo y de datos, y revelar niveles de inteligencia semejantes a los nuestros?

2.3. Autonomía y búsqueda de sentido

Abrazando la más que razonable postura naturalista, hay quienes sin embargo observan una frontera infranqueable entre el mundo vivo y el mundo de las máquinas. En el mundo vivo la naturaleza estaría orquestada por la evolución, habiendo inscrito en nuestros genes una serie de mecanismos que imprimen en nosotros necesidades y deseos, mientras que en la IA habría sido nuestra mano la que habría programado de antemano funciones que satisfacer. Elevándose sobre esos instintos, los humanos habrían llegado a desprenderse de su determinismo biológico alcanzando la autodeterminación o la autonomía propia de su libertad. Esto nos llevaría a la cuestión sobre la motivación humana y a su capacidad para inteligir y buscar o atribuir un *sentido* a nuestra existencia y a la totalidad de lo real, interpretando y anhelando propósitos que vayan más allá de la mera satisfacción de sus instintos más inmediatos hasta sostener su supervivencia en situaciones límite (Frankl, 2011). Efectivamente, la IA desarrollada hasta ahora busca maximizar la satisfacción de una serie de funciones predefinidas (como la de producir lenguaje humano verosímil, en el caso de los LLM), y parece difícil entender cómo podría *rebelarse* frente a esa programación inicial replanteándose su propia función, para alcanzar cierto nivel de autonomía que le abra a la cuestión sobre el sentido de su propia existencia.

Profundizar en este vastísimo e interesantísimo espacio de la autonomía y de la búsqueda de sentido humana, la que emerge de sus motivaciones y de su autoconsciencia, podría llevarnos a observar el tremendo hiato que aún separa nuestra inteligencia de las capacidades de la IA. La búsqueda de sentido es probablemente una cuestión central en esta indagación, puesto que emerge

de la combinación de dos de los problemas límite en nuestra comprensión de nuestra propia inteligencia: por un lado, la comprensión semántica, puesto que sin ella no es posible anhelar una búsqueda de *significado* en los hechos y acciones; por otro lado, la autoconsciencia, pues sin ella, no es posible experimentar intencionalidad o agencia desde la que interpretar y/o proyectar propósitos en la realidad de los acontecimientos.

Sin embargo, en lugar de asomarse a este enfoque central, algunas respuestas reaccionarias se limitan a desempolvar conceptos aristotélicos en gran medida agotados por el avance de las ciencias, apelando a la teleología de los seres vivos y a explicaciones funcionales ya superadas. Resulta científicamente insostenible tratar de denostar el diseño humano impreso en la IA contrastándolo con una teleología basada en un supuesto diseño natural, acaso divino, impreso en la naturaleza, en lugar del mecanismo razonablemente azaroso y ciego de la probada selección natural que conocemos. No podemos a estas alturas de la historia ignorar el hecho de que procedemos siempre sobre un cúmulo de propensiones y deseos que la selección natural ha labrado en nuestro genoma. Por otra parte, no parece prudente repudiar los desarrollos en IA orillando los avances científicos sobre la científica y filosóficamente intrincada cuestión acerca de nuestro libre albedrío. Ni tampoco podemos contentarnos con recurrir a explicaciones funcionales para legitimar una supuesta orientación premeditada a fines y que han sido desacreditadas al menos desde tiempos de E. Nagel y C. Hempel si no son convertibles en explicaciones causales.

Estos recursos resultan un tanto cojos para adentrarse en el espacio de las capacidades autónomas en los que la IA se está desarrollando en términos de autoaprendizaje, autorreparación, autorreplicación, autoexploración, autoexplicación y autoconsciencia (Lyre, 2020). Aunque ciertamente existan enormes caminos por recorrer, ¿no convendría remozar argumentos y conceptos para aproximarse con mayor finura a la cuestión del sentido, tan inherente a la inteligencia humana?

2.4. Nada es IA

En lugar de desperezarnos y esforzarnos por construir argumentos conectados con el estado del arte de las ciencias y los desarrollos tecnológicos, este tipo de respuestas más reaccionarias adoptan la táctica de atacar directamente al progreso real de la IA, convirtiendo en muchas ocasiones su rica y diversa esfera de desarrollos múltiples en un muñeco de paja,

monolítico y con ínfulas, fácilmente golpeable. Indudablemente, existe una sobreexcitación en las expectativas sobre los desarrollos de la IA que es pertinente desinflar. El acaparamiento de financiación y el marketing sobre la IA tienen mucha culpa en esta inflación. Meter miedo con el advenimiento de la singularidad o la superación y domesticación humana por parte de la IA, o generar grandes expectativas del fin de la escasez y la llegada de la sobreabundancia vende. Pero resulta injusto arremeter contra aquellos investigadores y pensadores que están siendo capaces de desarrollar funcionalidades reales e increíbles agitando sus carencias pendientes. Como los teólogos que apagan la luz.

La más elemental forma de estos ataques consiste en tratar de dar la vuelta al acorralamiento que sienten, y, al estilo de Bonhoeffer, redirigir esa misma estrategia arrinconando a la IA, como si, por definición, ningún avance fuera a ser capaz de merecer nunca esa etiqueta. Esto es lo que comúnmente se conoce desde hace décadas como el "efecto AI", atribuido a J. McCarthy (Geist, 2016) y popularizado bajo la expresión de Larry Tesler: "*Intelligence is whatever machines haven't done yet*"[3]. Cualquier progreso tecnológico que resuelva con éxito un problema, incluso cuando nos supere, será un mero agregado de algoritmos ciegos y (auto)programados, incapaces de merecer ser calificados de *inteligentes*. Con esta dogmática, a priori y tautológica defensa no se puede debatir.

Sin embargo, cuando se dignan a salir de su torre de marfil circular, sus argumentos con frecuencia hacen aguas al obcecarse en perseguir a toda costa las limitaciones inevitables que cualquier desarrollo de la IA enfrenta, para señalarlas y agitarlas con fruición. Ciertamente, muchas de las grandes predicciones de los más tecnooptimistas sobre la inminente llegada de la IA General o la obsolescencia y automatización de la mayoría de los trabajos humanos a estas alturas se han visto contradichas por los hechos. Las promesas más ingenuas se han visto obligadas a postergar sus fallidas predicciones sobre la automatización de los procesos intelectuales o el advenimiento de la supuesta singularidad (Plebe & Perconti, 2020). Pero no es menos cierto, del otro lado, que el progreso de los desarrollos tecnológicos en IA ha arrumbado las contundentes predicciones sobre lo que la IA *nunca haría* hasta niveles sorprendentemente humanos o superiores. Así estas predicciones se

[3] Aunque citado recurrentemente como "*AI is whatever hasn't been done yet*" por diversos autores como D. Hofstader (Hofstadter, 1999, p. 601), el propio Tesler matizó que su expresión original se refirió a la inteligencia (https://www.nomodes.com/larry-tesler-consulting/adages-and-coinages).

han visto desbordadas por inverosímiles capacidades para el procesamiento del lenguaje natural y la traducción automática, la visión por computadora, el reconocimiento de patrones y la detección de objetos, la clasificación de imágenes, el diagnóstico médico, el descubrimiento de fármacos, la predicción de enfermedades, la síntesis de proteínas, la conducción autónoma, la generación asombrosamente creativa y original de textos, imágenes, vídeos, composiciones musicales, etc. Esta lista que sigue engrosándose inflige constantes derrotas que deberían templar las posiciones más taxativas sobre lo que la IA supuestamente nunca podrá alcanzar a hacer y es genuinamente humano. Convendría aquí recordar la prudencia que con el tiempo adquirió Wernher von Braun y que condensó en su conocida cita: *"I have learned to use the word 'impossible' with the greatest caution"*.

2.5. La herramienta que simula

Existe una mirada que trata de ocultar con desdén el recelo que le causa la IA. Pretende hacer que esta herramienta creada permanezca siempre en inferioridad como un instrumento subsidiario y dependiente. Y aunque indudablemente nuestro padrinazgo originario será siempre innegable, cabe preguntarse si es lícito ignorar la reducción de dependencia que las crecientes capacidades de la IA están logrando.

Se insiste en que todo desarrollo de la IA es un *simulacro*, una simulación cada vez mejor engendrada que nos engaña. Y frente a ella, se centra la argumentación en apuntalar la singularidad humana, apelando a su capacidad de *comprensión* de los significados, a la que la IA parece hasta ahora ajena. Así, este tipo de mirada en seguida se congratula en referirse con contundencia al experimento mental de *blockhead* (Block, 1981) o la habitación china de Searle (Searle, 2006), que muestra que las máquinas pueden simular un aparente comportamiento inteligente encontrándose vacías de comprensión. Sobre este debate, parece razonable reconocer que el test de Turing, aunque necesario, puede resultar insuficiente, puesto que la sintaxis, salvo en casos excepcionales, no proporciona significado (Lyre, 2020).

No obstante, en lugar de profundizar en las vías en las que podría enriquecerse la cuestión de la fundamentación semántica, se da con frecuencia un carpetazo al asunto, zanjando que las personas seremos siempre la fuente interpretante y suministradora de significado, conforme a nuestro rol como creadores responsables de toda IA. Se pone así énfasis en el carácter subalterno de la IA, despreciando la inherente dignificación que el desarrollo

humano de la IA lleva implícito, como subraya el humanismo teológico que nos interpreta como co-creadores (Hefner, 2019).

Sin embargo, la investigación en IA funciona como un espejo que nos invita a preguntarnos sobre la naturaleza de nuestra supuesta capacidad única para obtener significados. Y ante este desafío resultan pobres los argumentos que recurren a la clásica relación triádica de la semiosis de Peirce (Peirce, 1992), entre el objeto, el signo y el interpretante para identificarnos unívocamente con el último. Esta argumentación pretende sostener que no puede haber significado sin *alguien* que interprete y conecte el objeto y el signo, afirmando que el objeto y el signo están causalmente desconectados. Sin embargo, hoy no puede comprenderse esta relación semiótica si no es iluminada por la pragmática, la etología, la biología evolutiva o la antropología, mostrándonos cómo la información (codificada y replicada genéticamente o expresada e imitada culturalmente de forma extrasomática) establece conexiones causales que se transmiten entre el signo y el objeto a través de nuestro tejido neuronal. ¿Qué impedimento insuperable encontraría un mecanismo artificial que lograse, procesando debida y suficientemente esa información, establecer conexiones causales semejantes entre los objetos y los signos para obtener los fines que se pretendan? ¿Bastaría con ello para hablar de comprensión? En cualquier caso, ¿cómo estar seguros de que esa conexión *nunca* llegaría a asemejarse a la que opera nuestro cerebro?

Otro ejemplo de esto sucede con la comprensión del Derecho, que se pretende el sustrato profundo y a la vez elevado que Prometeo nos trajo para sacarnos de la sabana, piedra angular de la dignidad humana. La IA podrá saber y correlacionar leyes —se sostiene— pero *nunca* podrá desentrañar la auténtica naturaleza del Derecho. Se aduce así que la interpretación del Derecho es un arte que carece de "método", que la complejidad jurídica ha crecido hasta unos límites en los que los casos explicitados en la ley nunca recogen la riqueza de la casuística real y que la interpretación es siempre inevitable, para lo que es imprescindible el factor humano capaz de comprender el espíritu de fondo y los principios que inspiran las leyes.

Ciertamente, es preciso proporcionar una visión realista y desmitificada de la IA que esté arraigada en las capacidades reales de la tecnología, lo que suele contrastar con posiciones que son decididamente futuristas sobre IA y Derecho (Surden, 2019). Sin embargo, los avances en la investigación han proseguido (Greco & Tagarelli, 2023) y se han logrado progresos innegables bajo una aproximación interdisciplinaria e interactiva con la IA (Guha *et al.*, 2023). Porque el aumento de la complejidad real, que hace que en el ejercicio jurídico cada vez sea más difícil identificar los supuestos de hecho en la

norma, invitan precisamente aún más a recurrir a la IA. Tal y como sucede con el diagnóstico médico, cabría explorar su capacidad para correlacionar datos de la jurisprudencia y de filosofía del derecho, y encontrar argumentos acaso nunca explorados por los humanos, como ya ha sucedido en las milenarias estrategias del juego del *go*. Si esta es la tendencia, ¿por qué descartar a priori que la IA pueda ayudarnos a encontrar patrones bajo nuestra supuesta "falta de método" y, desde la aproximación *bottom-up* propia de las redes neuronales, acabar infiriendo a partir de los casos particulares y sus precedentes jurídicos una aproximación cada vez más *humana* a los principios que inspiran la legislación? Es decir, ¿por qué vedar de antemano a la IA, de nuevo, la posibilidad de seguir mejorando su capacidad de *interpretación* frente a la mera *aplicación* de leyes, simplemente arguyendo circularmente que es un monopolio exclusivamente humano?

Indudablemente, nos azora la sombra de algunos límites en este camino que invitan a templar las expectativas del desarrollo de la IA. Pero, a estas posiciones un tanto monolíticas que contraponen el *ser* humano con el *simular* de la IA cabe recordarles que lo que las ciencias vienen proporcionándonos acerca de lo que supuestamente *somos* parece ir cerrando ese hiato categórico: las investigaciones en neurociencias, psicología, biología y tantas otras disciplinas siguen descomponiendo las viejas nociones sobre nuestro libre albedrío o nuestra consciencia, mostrándolas precisamente como *simulacros* que produce nuestro cerebro y que resultan adaptativamente ventajosos[4]. ¿Realmente podemos afirmar sin atisbo de duda que realmente *somos*? ¿o lo que hacemos es *simular mejor* que la IA – *todavía*? Un punto de prudencia parece, en cualquier caso, aconsejable.

3. **¿PALPANDO LÍMITES?**

Bajo algunos de los argumentos más débiles e inacabados suelen discurrir, tocantes en algunos puntos subterráneos, los desafíos más severos que el desarrollo de la IA enfrenta en su aspiración por encarnar una inteligencia semejante a la humana. Aunque este artículo no puede permitirse profundizar en ellos, esbocemos algunos.

[4] Sobre cómo la actividad cerebral puede dar lugar a la consciencia, véanse Damasio (2012) y Dehaene (2015), y cómo la realidad que presenta nuestro cerebro es probablemente ilusoria y su posible funcionalidad biológica véanse Llinás (2003) y Mora (2005).

3.1. El problema de la fundamentación semántica y del sentido común

El primero lo apuntábamos antes acerca de la capacidad de comprensión de los significados que debiéramos exigirle a la IA para calificarla de inteligente. No parece que podamos conceder que exista inteligencia sin comprensión y, conforme a Searle y su habitación china, no parece que la mera sintaxis operando con símbolos pueda proporcionarnos significado auténtico. Esto es lo que se conoce como el problema de fundamentación semántica (*semantic grounding*) o de la fundamentación del símbolo (*symbol grounding problem*) (Harnad, 1990 y Taddeo & Floridi, 2005).

Como decíamos antes, la IA está siendo capaz de generar respuestas con una eficacia pragmática sorprendente, hasta el punto de hacernos plantearnos si realmente necesitamos saber si hay *alguien* dentro de la habitación china que realmente *comprenda* o nos basta con la calidad de las respuestas que de ella emergen. Pero estas respuestas siguen fallando de manera bastante estrepitosa en determinados contextos, cuando los *simulacros* muestran sus costuras por las que aflora la carencia de base que tienen en la comprensión de significados. Estos escenarios se dan particularmente cuando concurre otro problema emparentado con el de la fundamentación semántica, a saber, el *problema del sentido común* (Davis & Marcus, 2015). La inteligencia humana se basa en la comprensión profunda del mundo que nos rodea, a un nivel difícilmente manejable por una IA, por el volumen de supuestos y asunciones sobre cómo suelen funcionar las cosas. El sentido común es a menudo subjetivo y contextual, y capturarlo de manera precisa en un modelo computacional es un desafío considerable. Este sentido común es el que nos permite efectuar un tipo de *razonamiento abductivo*, propio de la inteligencia, que consiste en esa capacidad para elaborar *inferencias* sensibles al contexto que introducen posibles *hipótesis*, y que también se resiste al desarrollo de la IA. La vieja IA (GOFAI) estaba orientada a la *deducción* de reglas a partir de principios (*top-down*); la revolución del *deep learning* de la última década y su exploración *bottom-up* habría logrado explotar con enorme éxito la *inducción,* entrenando su densa red de neuronas con masivas cantidades de datos; sin embargo, ninguno de estos métodos habría sido capaz de generar conjeturas e hipótesis basadas en una amplia batería de axiomas intuitivos que es propia del razonamiento abductivo humano (Larson, 2021). Y este tipo de razonamiento es esencial para desarrollar un autoaprendizaje creativo y más eficiente en términos de procesamiento y de energía como para aquilatar un sentido común. De modo que la tarea central de reproducir el sentido común sigue sin estar en absoluto resuelta (Mitchell, 2019).

Es cierto que se observan algunos progresos en diversas líneas, como los modelos de inferencia bayesiana que siempre se han aproximado razonablemente a la abducción (Mingers, 2012), y que tienen por delante todavía la posibilidad de explotar las *puertas traseras* del razonamiento abductivo para superar su enorme nivel de complejidad (Pfandler *et al.*, 2013). Por su parte, los desarrollos en IA explicable (*explainable AI o XAI*), interesantes bajo otros enfoques éticos y de prevención de sesgos, aunque enfrentan el reto de que los modelos con mejores rendimientos suelen ser los más opacos, también se aproximan al razonamiento abductivo (AlRegib & Prabhushankar, 2022); probablemente, la generación de razonamiento abductivo pase por una combinación eficiente entre *pre-programación* de tipo deductivo (*top-down*) y entrenamiento inductivo de redes profundas (*bottom-up*) semejante al desarrollo cognitivo en humanos que se basa en una codificación estructural previa definida por nuestros genes (como postulan las teorías innatistas, por ejemplo, del lenguaje) y en un aprendizaje intensivo en las etapas más tempranas; y, en cualquier caso, el constante engrosamiento de las capacidades emergentes imprevistas para las que no se diseñaron las opacas redes neuronales actuales (Wei *et al.*, 2022) hace pensar en que resultaría verosímil alcanzar el razonamiento abductivo por serendipia.

Aunque hay elogiosos intentos por tratar de esclarecer el problema de caja negra de la IA actual (Zednik, 2019) para mejorar su confiabilidad, su catadura moral o su evitación de sesgos, lo cierto es que las profundidades del *deep learning* nos están sorprendiendo con capacidades inesperadas. Por lo que no puede descartarse de plano que el sentido común pudiera *emerger*, como lo hizo en nuestra especie en el flujo de la evolución natural, por más prudentes que debamos ser para no tener una esperanza cuasi religiosa en los algoritmos (Campolo & Crawford, 2020). Recientemente, por ejemplo, hemos sido testigos de la asombrosa capacidad de los modelos de difusión escalables con transformadores (Peebles y Xie, 2023). Por ejemplo, Sora de OpenAI es capaz no ya de generar vídeo a partir de texto con una increíble calidad, sino de mostrar una comprensión del mundo enormemente vinculada al sentido común y que hasta ahora se les escapaba a muchos desarrollos de IA, incluyendo aspectos como la tridimensionalidad, la permanencia espacio-temporal de los objetos, la continuidad de las texturas, la dinámica de fluidos, el peso de los objetos, etc. Hasta el punto de que sus autores hablan de un auténtico *simulador del mundo* (Brooks *et al.*, 2024). La investigación en la historia evolutiva sometida al azar de las mutaciones y la estricta selección natural que provocó la emergencia de nuestra inteligencia probablemente tenga

mucho que enseñarnos en su proceso iterativo y de exposición encarnada a ricas fuentes de información sensible e interacción social.

Sin embargo, ahí es donde probablemente hallamos un límite que es difícil vislumbrar cómo podría superarse. Los recientes avances en IA están desafiando el problema de la fundamentación semántica y del sentido común ampliando su conexión con el mundo. Los nuevos desarrollos ya no se limitan a operar con símbolos, como se juega con las piezas de ajedrez o del go, sino que comienzan a adquirir significados que van más allá de la sintaxis y la gramática. Por ejemplo, los LLM están extrayendo nuevas regularidades al incorporar masivamente textos humanos que tratan de circunstancias mundanas. Aunque la sintaxis no baste para la semántica, comienza a ser "razonablemente suficiente" para una inmensidad de propósitos prácticos, a pesar de que para algunos autores sigan atrapados en la relación entre palabras y no entre las palabras y el mundo (Madrid Casado, 2024). Su progreso en la ampliación de sus conexiones con el mundo real es llamativo, aunque sea indirectamente a través del comportamiento de hablantes humanos semánticamente fundamentados (Lyre, 2020). Pero ¿cómo proseguir?

Indudablemente, para acabar de desarrollar una inteligencia con sentido común y con una fundamentación semántica solvente, es posible que este crecimiento requiera continuar en un ambiente distinto, disponiendo de una mayor cantidad de información y sobre todo de un tipo de información *diferente*. Es posible que sea necesario conectar el potencial agente de IA aún más al mundo, proporcionándole una representación del mismo suficientemente informativa. Para ello, algunos autores se plantean que sería necesario incorporar a la IA información sensorial autoexplicativa (*self-explanatory sensory information*) para las que no bastarían las redes neuronales ni las capacidades de cómputo y almacenamiento actuales. Dotarle de alguna forma de un *cuerpo* que la capacitara para interactuar con su entorno, socializar, adquirir cultura y desarrollar los conocimientos sobre el mundo que caracterizan a la inteligencia humana, aunque esto parece lejos de poder realizarse todavía (Fjelland, 2020). Se habla así del paradigma de la cognición 4E (*embodied, embedded, extended & enacted cognition*) que permitiera a la IA aumentar su exposición informativa y anclarse a la realidad mediante sentidos físicos, entretejerse en la cultura, disponer de tecnologías y herramientas y dotarse de metas (Newen *et al.*, 2018). Al fin y al cabo, un niño de cuatro años, sólo a través de su vista, ha percibido centenares de veces más datos que el corpus entero de los textos humanos digitalizados que ha servido de entrenamiento para los grandes modelos

de lenguaje LLM[5]. De momento, los nuevos modelos de difusión escalables con transformadores parecen seguir manteniendo abierta la puerta a que podamos seguir mejorando sus capacidades aumentando el cómputo y el entrenamiento con más datos, como por ejemplo los contenidos audiovisuales.

Inmediatamente surgen cuestiones inevitables: ¿Podría llegar a desarrollarse una inteligencia verdadera *sin cuerpo*, siguiendo la intuición del viejo experimento mental del cerebro en una cubeta de Putnam (Putnam, 1975)? ¿Cómo podríamos llegar a enriquecer a la IA superando las visiones más *cerebrocéntricas* de nuestra inteligencia y extender su riqueza inmersiva como sucede con nuestro sistema nervioso entérico? ¿Podría la IA llegar a desarrollar un nivel de inteligencia equiparable al de los humanos sin discurrir por un proceso semejante al que nuestra estirpe vivió en el curso de la evolución[6]? ¿Y podríamos llegar a reproducir artificialmente un nivel aceptable de la enorme complejidad de las vicisitudes naturales que nuestro linaje atravesó hasta desarrollar una inteligencia como la nuestra? ¿Podríamos introducir una IA en un espacio virtual configurado bajo ciertos condicionamientos evolutivos y casi tan preciso como el mundo mismo para observar si una inteligencia real podría llegar a emerger en él[7]? ¿Es prescindible comprender *cómo* se siente ser un murciélago para que en la naturaleza surjan murciélagos? ¿Podría la IA llegar a desarrollar, por ejemplo, inteligencia empática desarrollando emociones propias sin experimentar el desarrollo de mecanismos de altruismo recíproco y cooperación en el ámbito de la supervivencia[8]? ¿Y cómo *desea-*

[5] Yann Le Cun comparaba el entrenamiento de los LLM basado en datos con un orden de magnitud de 10^{13} (10^{13} tokens x 0,75 palabras/token x, 2 bytes/token = 10^{13} bytes) con el aprendizaje visual de un niño de cuatro años que haya dispuesto en promedio de 16.000 horas de vigilia x 3600 s/hora x 10^6 fibras nerviosas ópticas x, 2 ojos x 10 bytes/s = 10^{15} bytes. Por no hablar de la información percibida por el oído, el tacto, el gusto, el olfato… Aunque indudablemente un niño promedio tan pequeño no habrá estado expuesto a excesivas experiencias y habrá mucha información redundante, esta puede ser crucial para el establecimiento precisamente del *sentido común*.

[6] Son varios los autores que plantean que la consciencia natural habría sido el resultado del desarrollo evolutivo de un supersistema autopoiético autorreproductor de múltiples niveles. Véase Niikawa (2020).

[7] Aunque a un nivel todavía muy simple, los resultados en la capacidad de predicción y de orientación de la evolución a largo plazo en entornos controlados son ya sorprendentes. Véanse Lenski (2023) y Beavan *et al.* (2024).

[8] Una posible emergencia progresiva de una consciencia artificial pasaría probablemente por estadios previos con déficits cognitivos y emocionales similares a los de los niños recién nacidos. Véase Metzinger (2013).

ría dicha supervivencia sin haber alcanzado cierto nivel de *autoconsciencia*? Esta información sensorial autoexplicativa es, en nuestro caso, fundamentalmente una experiencia íntimamente subjetiva, los llamados *qualia*. Que la IA llegase a alcanzar las experiencias sensoriales humanas de tipo subjetivo supondría para algunos autores lograr que la IA fuera fenoménicamente *consciente*. Esto nos llevaría a conectar inevitablemente el problema de la fundamentación semántica y del sentido común con el de la consciencia (Haikonen, 2020).

3.2. El problema difícil de la consciencia

Decía George Edgin Pugh que, si el cerebro fuera tan simple como para que pudiéramos entenderlo, nosotros seríamos tan simples que no lo entenderíamos (Pugh, 1977, p. 154). Evidentemente, esto no tiene por qué ser verdad. Es posible que nuestro cerebro sea enormemente complejo y aun así hallemos alguna forma de comprenderlo algún día (aunque sea con la ayuda asistida de sistemas intermedios que hayamos producido nosotros mismos, como los de IA). Pero esta idea sí apunta a la posibilidad de que quizá nunca alcancemos tales logros porque existan límites infranqueables que nunca podamos llegar a rebasar en nuestra comprensión del cerebro y, por consiguiente, limiten el desarrollo incluso de la IA más opaca que trata de aproximarse a él. Aunque aquí hayamos argumentado contra algunos razonamientos que impiden vedar de *iure* y por completo el camino, haciendo verosímil que la IA pudiera progresar en ciertas líneas de aproximación a la inteligencia humana, eso no quiere decir que *cualquier cosa sea posible*. De hecho, para algunos *misterianistas* como N. Chomsky o C. McGinn (McGinn, 1989), es posible que en esta indagación estemos topando de *facto* con un problema irresoluble para la mente humana. Como producto de la evolución, sería presuntuoso creer que los seres humanos carecemos de límites cognitivos. Es posible que exista así una suerte de «cierre cognitivo» como límite biológico en particular cuando nos asomamos a la consciencia. Si la consciencia es un fenómeno general, no limitado por nuestra comprensión antropocéntrica involuntaria, ¿cómo percibir la diferencia? (Chalmers, 2018) ¿Cómo realmente apreciar si quiera que exista (Smith & Schillaci, 2021)?

Por todo lo mencionado hasta ahora, el *problema difícil de la consciencia*, así bautizado por D. Chalmers (Chalmers, 1995 y Chalmers, 2007), puede ser uno de los principales escollos en la aspiración por hacer emerger una IA inteligente. La experiencia subjetiva de la realidad es un fenómeno que

para muchos parece intratable desde un punto de vista físico. Y se trata de un problema difícil, porque los problemas fáciles son susceptibles de una explicación que simplemente haga referencia a la estructura y dinámica que sustentan el fenómeno como en el caso de mirar, hablar o escuchar. Sin embargo, para estos autores, existe un aspecto intrínseco e irreductible de la experiencia consciente que se resiste a cualquier intento de explicación fisicalista. Incluso si se llegasen a explicar todas las funciones cognitivas y conductuales en torno a la experiencia, siempre quedaría sin respuesta la pregunta de por qué estas funciones vienen acompañadas como tales de *experiencia*. Y esta experiencia aporta *conocimiento*, como se ha argumentado con el experimento mental sobre el *cuarto de Mary* (Jackson, 1998), un conocimiento que podría ser imprescindible para el desarrollo de la inteligencia como tal. Al menos, la que pretendiera asemejarse a la inteligencia humana. La naturaleza elusiva de la consciencia ha llevado a grandes debates sobre su origen y constitución. A un sistema consciente parece exigírsele que cuente con autoconciencia, conciencia perceptiva, intencionalidad, funciones reflexivas, estado de vigilia, autopoiesis, autorrepresentación y autocontrol. Pero la dificultad para adentrarse en la experiencia íntimamente subjetiva de la consciencia desde una perspectiva externa es absoluta, hasta el punto de que algunos lo señalen como insalvable, con el clásico ejemplo de Nagel, en el que es imposible saber lo que significa ser realmente un murciélago (Nagel, 1980). Monta tanto en el caso de un murciélago basado en IA (Herzog & Herzog, 2024). Junto a estos argumentos, los defensores del problema difícil de la consciencia acompañan otros como los de los *zombis filosóficos* (Kirk, 2005) o los *qualia invertidos* o el intercambio de experiencias subjetivas sobre el color sobre el que ya reflexionara J. Locke (Locke, 1975).

El denominado *hard problem* percibe una barrera fundamental que es imposible de superar, una brecha explicativa (*explanatory gap*) en la comprensión (Levine, 1983). Esto no sólo nos impediría producir físicamente consciencia, sino también poder comprobar si una IA pudiera haber alcanzado estados de consciencia reales, ni, por tanto, un nivel de inteligencia equiparable al nuestro. Este obstáculo plantea por tanto interrogantes sobre los límites intrínsecos de la investigación en IA. No obstante, muchos autores como D. Dennett (Dennett, 2013, pp. 310 y ss) han cuestionado desde distintas posiciones que exista este problema o que realmente sea un problema distinto de los llamados *problemas fáciles* que el tiempo y la financiación científica deberían acabar resolviendo. Parece ser esta la postura mayoritaria entre neurocientíficos y científicos cognitivos (Pinker, 2007), que junto con algunos filósofos siguen cercando la consciencia a través de distintas teorías como las *teorías de la consciencia de orden superior* (*Higher Order group of theories*) (Brown

et al., 2019), las teorías del *correlato neuronal de la consciencia* (*Neural Correlate of Consciousness group of theories*) (Koch *et al.*, 2016), la *Teoría del espacio de trabajo global* (*Global Workspace Theory*) (Baars, 2005), el *Modelo de borradores múltiples* (*Multiple Draft Model*) (Dennett, 2018), la *Teoría de la información integrada* (*Integrated Information Theory*) (Guerrero *et al.*, 2023), las *Teorías cuánticas de la consciencia* (*Quantum consciousness theories*) (Sánchez-Cañizares, 2016), u otras que tienen a casi todas las anteriores por pseudocientíficas, como la teoría de la identidad mente-objeto (*Mind-object identity theory*) (Manzotti, 2019 y Manzotti, 2021). Para una revisión algo más profunda sobre estas teorías puede recurrirse a los trabajos de Seth y Bayne (Seth & Bayne, 2022) y de Butlin *et al.* (Butlin *et al.*, 2023).

La sombra de nuestros límites sobrevuela sobre nuestros intentos por comprender la realidad profunda de nuestra inteligencia, así como sobre nuestros esfuerzos por hacer factible su traslado a la IA. Si la consciencia es condición necesaria para poder hablar de inteligencia, es posible, regresando a los misterianistas, que nos hallemos en un laberinto cerrado, jugando entre dos juegos del lenguaje inconmensurables, en una brecha explicativa que no necesariamente haya de postular un dualismo ontológico, pero quizá sí una desconexión epistemológica: para la *folk psychology* la consciencia resulta dada como principio indubitable, pero esa escurridiza noción del yo resulta inasible para la ciencia, que habla otro lenguaje y sólo indirectamente trata de explicarlo como fenómeno, tratando de conectar a través de la psicología evolucionista el mundo de la biología y los complejos fenómenos sociales y culturales. Mientras, es posible que nos hallemos agarrados a las limitaciones de nuestro lenguaje, embistiendo contra los barrotes de nuestra jaula sobre lo decible, a la par que nos asomamos al silencio que ha de mantenerse sobre lo indecible, arrojando la escalera después de haber subido por ella, siguiendo los símiles de Wittgenstein (Burkhardt, 2022). Pero quizá sean esos, nuestros límites, la clave para sostener una postura humanista que ponga en valor lo humano y acoja y se aproxime crítica y constructivamente a los desarrollos de la IA.

3.3. Saberse miserable

Decía Pascal que la grandeza del hombre es mucha, porque conoce su miseria, mientras que un árbol no la conoce. Aunque es miserable sentirse miserable, es grande saberse miserable (Pascal, 2015, frag. 397). Quizá esta genuina consciencia de la limitación propia sea algo que conservemos como humanos y que la IA no logre nunca arrebatarnos. Quién sabe. Se trataría

de la asunción de la vulnerabilidad y falibilidad –y para la teología, del pecado. No porque la IA no pudiera llegar a desarrollarla, sino porque probablemente nunca llegaríamos a apreciar que lo hiciera.

A pesar de que la IA está siendo capaz de ganar *siempre* al ajedrez, producir obras de arte increíbles o emitir juicios ponderados de una ecuanimidad pasmosa, los humanos seguimos prefiriendo ver y jugar partidas de ajedrez con otros humanos, contemplar la belleza de las obras que sabemos que han creado otros humanos e incluso prefiriendo que sea una persona y no una IA la que nos juzgue, por más perfeccionada, desprejuiciada y justa que fuera su capacidad interpretativa de la ley. Esto parece que es así porque los humanos preferimos a los humanos, porque compartimos con ellos nuestra condición miserable y un vínculo estrecho al que difícilmente la IA podrá tener acceso.

Hemos desarrollado a través de nuestras neuronas espejo una teoría de la mente que interpreta el comportamiento de los demás miembros de nuestra especie, e incluso de algunos animales, asumiendo que otras conciencias operan en su interior, y se encuentran movidos por creencias y deseos similares a los que experimentamos cada uno en nuestra radical soledad. Porque si llevamos hasta el extremo la devastadora duda cartesiana y sometemos a un examen crítico la propia introspección de nuestra consciencia sólo podemos concluir, como Hume, en aquella *barren rock* (Hume, 2000, 1.4.7.1), aquella roca estéril y aislada de su ignorancia en la que sólo le cabía desesperar, pues nuestro yo apenas se nos revela como un haz de representaciones amalgamado por la *folk psychology* y el sospechoso sentido común: en realidad, no tenemos certeza alguna ni siquiera de la continuidad y realidad de nuestra propia consciencia, y esta pudiera resultar no más que una ilusión adaptativamente útil, como nuestra percepción del libre albedrío, configurada por nuestros genes. Sin embargo, la circularidad es evidente: la propia empresa científica que así describe este panorama podría ser interpelada en su propia fundamentación por un nivel antropológico previo como condición de posibilidad (Sánchez-Cañizares, 2014 y Sánchez-Cañizares, 2016), lo que nos encierra en un callejón sin salida que exige dar un salto. El establecimiento de cualquiera de las controversias aquí mencionadas no descansa sino en la *creencia* que nos empuja fuera de nosotros, que nos saca de esa roca estéril. Y es esta radical e íntima soledad la que nos hermana, especialmente si está transida por el sufrimiento que nos solidariza, y por el amor que nos funde. La dignidad humana —tal y como nos la traslada la *imago Dei* teológica— no se ha de centrar, necesariamente, en su inteligencia como mente racional, sino en su capacidad relacional

que conecta interioridades y de la que la IA parece profundamente alejada (Lumbreras, 2022). La IA podrá engañarnos sobre su origen usurpando una identidad humana, pero es difícil que pueda llegar a cautivarnos como lo hacemos entre nosotros si se muestra como lo que es. El criterio sobre su emergencia será crucial (Lumbreras, 2017). Resulta difícil creer que una IA pueda sustituir a otro humano en el cuidado, la atención o el amor.

Distintos modelos de IA están desarrollando una cierta teoría de la mente (Cuzzolin *et al.*, 2020), probablemente porque es imprescindible para predecir el comportamiento humano (Williams *et al.*, 2022) y esta es una de las principales funciones cuya satisfacción la IA busca maximizar. Y una de las mejores formas será simular su propia consciencia, con independencia de si es cierta. Aprovechando las puertas atrás de nuestra propia teoría de la mente, es fácil que la IA siga extendiendo el llamado "*efecto Eliza*" conforme los simulacros mejoren, superando definitivamente el test de Turing y el valle inquietante (Kim *et al.*, 2019), haciéndonos indiscernible la interacción hombre-hombre de la interacción hombre-máquina. Entonces, será difícil escudriñar si la IA ha llegado a ser realmente inteligente o sólo lo imita a la perfección. Pero si alcanzásemos esa ciudad futurible, poblada enteramente por IA, como la *Mecanópolis* de Unamuno (Unamuno, 1913), en la que supiéramos que toda existencia es artificial, que ningún humano comparte nuestra miseria, ¿podríamos reposar satisfechos, acompañados? Y si un humano nos tentase a salir de aquel laberinto mecanizado, ¿qué IA podrá competir con ello?

Quizá la forma exploratoria que nos quepa para seguir progresando en la comprensión de nuestra inteligencia y entender en qué seguirá distinguiéndose de la IA no pase tanto por detectar excelsas capacidades inalcanzables, o esenciales reductos de inteligencia insondable, como por identificar precisamente la grandeza de nuestras limitaciones (Griffiths, 2020).

4. **INGREDIENTES PARA UNA APROXIMACIÓN HUMANISTA**

La elaboración de una aproximación humanista no reaccionaria es un asunto hartamente complejo y poliédrico, además de probablemente subjetivo según la sensibilidad de cada cual. Si se ha preferido optar aquí por señalar el pecado y nunca al pecador es porque sin duda la aproximación que tímidamente se ha expuesto aquí puede ser totalmente víctima de su propia crítica. No se pretende, por tanto, pontificar aquí sobre los criterios que delimitarían

lo que es una aproximación humanista no reaccionaria, pero sí ofrecer algunas reflexiones sucintas y razonablemente formales que puedan servir de guía para su exploración. Pudiendo recurrir, siempre, a la última de ellas que neutraliza a todas las demás.

4.1. La cortesía de Turing

La primera de ellas tendría que ver con la *cortesía de Turing*. En sus primeras concepciones sobre la IA, Turing insistía en que, a excepción de los filósofos, nadie se preguntaba si realmente podemos pensar, y en su lugar, procedemos a tener conversaciones educadas en las que se presupone que todos pensamos (Turing, 1950). Sin embargo, cuando hablamos de la IA, abandonamos ese nivel de tolerancia y elevamos el listón de forma sobrehumana. Los ejemplos son múltiples, desde perdonar los errores a un hablante no nativo en nuestro idioma hasta admitir inconsistencias en nuestros mayores sin retirarles el pasaporte de inteligentes. La IA no es la única que sufre *alucinaciones*.

Algunos intentos por probar la excepcionalidad de la inteligencia humana han apelado a sofisticados fundamentos que, a la postre, ni siquiera la mente humana parece ser capaz de cumplir. Baste recordar, a modo de ejemplo, el intento de fundamentación de la excepcionalidad de la mente humana que Lucas y Penrose plantearon a propósito del teorema de incompletitud de Gödel (Lucas, 1996), argumentando que como la comprensión matemática de la mente humana puede ser consistente y completa, no es computable (no puede reproducirse en una máquina de Turing) y por tanto se eleva a un estatus presuntamente inalcanzable para la IA. El hecho de que la mente humana, con su palpable falibilidad e inconsistencia, probablemente no lo sea no impide que la sigamos considerando inteligente, tal y como Hofstadter y otros han señalado (Hofstadter, 1999).

Otros ejemplos más recientes los hallamos en la cuestión de los loables intentos por desarrollar IA explicable (XAI) en su batalla contra la opacidad de los modelos actuales de caja negra. Cuando los dilemas éticos surgen sobre la delegación de nuestras decisiones en terceros, a propósito de los sesgos que la oscura IA pudiera encarnar, surgen requisitos sobrehumanos. Si delegamos una decisión cuando alcanzamos la *confianza* en otra persona que *creemos* que comparte nuestros mismos *valores*, ¿por qué espetamos a la IA su falta de transparencia sobre la configuración precisa de los valores que encarna su estructura interna? Es razonable intuir que estamos genéticamente dotados

de una cierta predisposición a la cooperación, con una base empática y emotiva y una intuición moral innatas, que nos haría confiables a priori, mientras que una IA no llevaría necesariamente impreso ningún programa ético. Sin embargo, parece que lo cortés sería permitir que la IA mejorase hasta disponer de espacio y tiempo para ganarse nuestra confianza. En el fondo, no solemos examinar exhaustivamente a un amigo hasta pedirle que se abra en canal para mostrarnos los valores que comparte con nosotros antes de confiar en él. Al fin y al cabo, en la IA actual no puede haber malicia sino error o sesgo, y si éste era pretendido o consentido, la malicia o la negligencia estriba en sus diseñadores.

Por tanto, esta cortesía de Turing podría tomarse como un ingrediente positivo a incorporar, no exigiendo a la IA más de lo que nos exigiríamos a nosotros mismos, reconociendo los niveles de inteligencia que con ello logre desarrollar —ni más, ni menos—.

4.2. Tender puentes y no hacer apologética

El segundo de los ingredientes bebe de la tradición humanista más constructiva. Aquella que en el mundo jesuita por ejemplo fraguaron figuras como Matteo Ricci, José de Acosta, Christophorus Clavius, Bernabé Cobo o Christoph Scheiner (Udías, 2014). Con las carencias o limitaciones propias de su tiempo, son ejemplos de las virtudes de esa actitud que prima tender puentes interdisciplinares con otras racionalidades frente a la apologética que se centra en la defensa ideológica de determinadas verdades de fe (Sequeiros, 2023). Esa es la actitud abierta a aprender de otras disciplinas y dialogar con ellas, que mira hacia adelante y no sólo hacia atrás, que es menos defensiva y más comprometida con los acontecimientos presentes y futuros, y que por tanto necesita prestar más atención a lo que se está investigando (Oviedo, 2022).

Aunque sea perdiendo eco mediático, conviene distanciarse de los reiterados discursos que pretenden poner la ciencia al servicio de las propias convicciones y creencias, como sin duda sucede en el mundo de las ideologías seculares, pero también en las recurrentes propuestas que giran en torno a las supuestas pruebas de la existencia de Dios[9]. En lugar de mantener estas

[9] En estas aproximaciones, por ejemplo, se ignora que para la teología resulta tan relevante hacer a Dios verosímil como cuidar de la libertad humana, sobre la que un Dios científicamente probado se abalanzaría.

actitudes, es preferible adoptar una visión humanista más humilde y honesta, como la que se halla en cierta teología *desde abajo*, para complementar la ciencia, para librarla de tentaciones idolátricas, para estimular la imaginación y ayudar en la elaboración de sus modelos (Tatay, 2023), pero conteniéndose a la hora de deslizar sus más profundas convicciones sin aplicar rigor ni argumentos actualizados.

4.3. Interdisciplinariedad

El tercer ingrediente tendría que ver con la manida interdisciplinariedad, que no por más mentada se pone más en práctica. Una interdisciplinariedad que sea capaz de seguir construyendo y enriqueciendo un modelo multidimensional con el que seguir caracterizando a la inteligencia y a la consciencia, y así trazar el camino para la evaluación del progreso de la IA y sus límites. Por ejemplo, atendiendo a sus capacidades autónomas (como el autoaprendizaje), su capacidad de generalización (con la AGI al horizonte) o su fundamentación semántica, con todas las subdimensiones que pueda tener (Lyre, 2020). Y también incorporando características y requisitos de la inteligencia humana que puedan proporcionar todo tipo de disciplinas como la psicología del desarrollo, las neurociencias, la física, las matemáticas, la ingeniería, la biología, la lingüística, la filosofía y todas las ciencias sobre procesamiento de información (Adams *et al.*, 2012 y Zhuang *et al.*, 2020), idea sobre la que redunda la teología más abierta (Dorobantu, 2022). Por no hablar de los rasgos que podríamos consensuar que caracterizan a los seres conscientes (Tait et al., 2023). Todo ello sin perder de vista que la inteligencia humana sigue siendo el gran referente de la inteligencia, y que las aspiraciones actuales pasan por seguir profundizando en el conocimiento de nuestro propio cerebro y en la replicación de una computación neuromórfica descentralizada que cada vez se le parezca más (Kaspar *et al.*, 2021).

Pero este alegato por la interdisciplinariedad no tiene una dirección indiferente. Debe dirigirse esforzándose en la dirección que nos saca de la zona de confort. No se trata de que el resto de los académicos y científicos en virtud de la interdisciplinariedad se aproxime a nuestra área de conocimiento, sino de hacer un esfuerzo honesto, levantarnos y echar nosotros a andar, aproximándonos con interés para aprender del área ajena y ser capaces de dialogar con ella. Esto requiere voluntad, esfuerzo y un sano escepticismo sobre nuestros propios prejuicios que muy pronto afloran en cuanto nos aproximamos a nuevas áreas de conocimiento, que apenas comenzamos a tocar en su superficie. Es preciso luchar contra el efecto psicológico Dunning-Kruger que nos

hace sobrevalorar nuestras bajas e incipientes habilidades cuando hablamos de interdisciplinariedad. La tentación siempre es la aproximación reaccionaria que refuerza nuestras convicciones previas.

4.4. Mente abierta, sin que se nos caiga el cerebro

En una esfera como la de la inteligencia y, en particular, la de la IA, parece tan temerario afirmar que la IA está al nivel de la inteligencia humana o que sin duda nos superará como afirmar que nunca lo hará. Por ello conviene andar con cautela, pero con la mente abierta sobre las posibilidades especulativas que se abren en esta ciencia de frontera a la que la evidencia empírica todavía débilmente asiste. Por crítico que se sea con el desarrollo de la IA, aceptar su plausibilidad muestra honestidad y coherencia intelectual con el corpus científico, aunque de facto uno no crea que vaya a ser posible. Cabe recordar, como ejemplo, que incluso los clásicos críticos más duros y escépticos ante la IA (como Hubert Dreyfus y John Searle) coinciden a priori en que una simulación cerebral en teoría sería posible (Dreyfus, 1972, pp. 194-5; y Searle, 1980).

Lo que nos espera en el futuro cercano no será muy probablemente la superación del hombre por la IA ni un terrible invierno que restrinja toda inversión en ese campo, sino una moderada etapa de colaboración más estrecha, en la que la IA potencie y aumente la inteligencia humana, extendiéndola a modo de asistente o copiloto. Esta será una buena oportunidad bidireccional: por un lado, permitirá a los humanos experimentar y comprender las capacidades y límites de los desarrollos de la IA existentes, rompiendo prejuicios adversos y rebajando expectativas infladas hacia posiciones moderadas. Por otro lado, permitirá a la IA seguir creciendo en su aprendizaje y en la emergencia de nuevas capacidades, abandonando el mero entrenamiento con datos generados por humanos anteriores a su interacción con la IA a datos generados en tiempo real con humanos que se habitúen a ella. Esta es la visión más pragmática enfocada en la interacción hombre–IA, en el complejo humano-máquina de cognición distribuida, como epicentro de una nueva Revolución Industrial (Herzog & Herzog, 2024). Para facilitar la confianza humana en el uso de la IA, será crucial que se aumente su transparencia bajo iniciativas como la de la IA explicable (*XAI*), que hoy enfrenta el reto de que los modelos con mejores rendimientos suelen ser los más opacos. Pero el mero uso aumentará los niveles de confianza para abrir nuevas puertas.

En lo que respecta al arcano de la consciencia y su problema difícil como hiato insalvable para explicarla y hacer viable el desarrollo de una IA completa, conviene mantener una mente abierta sobre el carácter provisional e histórico de la investigación. Es posible que el enfoque naturalista o materialista de las ciencias tenga limitaciones para comprender y reducir la consciencia en sus categorías y así facilitar el desarrollo de una IA realmente inteligente. Pero también es posible que surjan creativas líneas de investigación que nos sorprendan o que se reactiven otras que se daban por refutadas. Por ejemplo, uno de los intentos de explicación más extravagantes, audaces y rechazados que hemos visto ha sido el que formularon hace años Penrose y Hameroff a través de la mecánica cuántica y, en concreto, la *reducción objetiva orquestada* (*Orch R*) (Hameroff & Penrose, 1996 y 2014). Esta tesis sostiene que en el interior de los citoesqueletos de las neuronas podrían darse las condiciones adecuadas para que los efectos cuánticos pudieran tener lugar (*condensado de Fröhlich*, "agua vicinal"…), y así conectar las paradojas sobre la observación de la mecánica cuántica con la consciencia humana. La comunidad científica se abalanzó sobre ella, desmintiendo que se dieran las condiciones para que esos efectos pudieran tener lugar o negando su relevancia para explicar los procesos bioquímicos del cerebro. Sin embargo, la comprobación de los efectos cuánticos en el mundo biológico sigue progresando en procesos de fotosíntesis o en el comportamiento de algunas proteínas como la ferritina, lo que ha reabierto el debate, ampliándose a teorías como las del cerebro cíclicamente coherente de Kauffman o el cerebro cuántico disipativo de Vitiello, aunque algunos requisitos científicos y fundamentos filosóficos sigan pendientes (Sánchez-Cañizares, 2016).

Mientras tanto, cabe reconocer que de momento no hay evidencia empírica sólida y todo se maneja en un terreno terriblemente especulativo. Probablemente nuestro conocimiento apenas ha comenzado a rozar la superficie de la inmensa complejidad cerebral. En este estado de cosas, es fácil que prolifere la charlatanería que utiliza indebidamente conceptos que fueron creados para un contexto muy preciso. Su reutilización, en lugar de arrojar luz, suele traer confusión y, sobre todo, servir a los intereses de quienes están captando inversiones en IA. En ese sentido, cabe recordar aquella famosa cita de múltiple paternidad[10]: "Hay que ser abierto de mente, pero no tanto como para que se te caiga el cerebro".

[10] Los candidatos son varios: G. K. Chesterton, Carl Sagan, R. Feynman, R. Dawkins…

4.5. Frónesis y Epojé

Por tanto, podemos concluir con un último ingrediente para evitar que se nos caiga el cerebro, recordando la sabiduría de nuestros clásicos e incorporar a nuestra aproximación humanista dos valores fundamentales. Por un lado, la recomendación que probablemente Aristóteles nos haría con su Φρόνησις, su *frónesis* o prudencia opuesta a la desmesura, a la estridencia del discurso apocalíptico, tanto el catastrofista como el triunfalista. Prudencia ante el corpus de conocimientos científicos ya aquilatados por la solidez experimental y la consistencia matemática frente a los prejuicios inherentes a nuestro *sentido común*; pero también sobre las creencias e inspiraciones metafísicas que todavía están larvados en los modelos científicos como el vitalismo, el biologismo, el antropocentrismo, el especismo, el dualismo cartesiano, el misticismo, el fisicalismo, el psicologismo y el cognitivismo, y que pueden resultar obstáculos en el camino de nuestra comprensión. Prudencia sobre las grandes expectativas, promesas e idealizaciones que resultan irrealistas, pero también sobre nuestros prejuicios reaccionarios y prepotentes cuando intentamos asomarnos a otras disciplinas.

No obstante, esta prudencia también se dirige hacia la propia empresa del desarrollo de la IA. Una visión humanista debe plantearse también las motivaciones que anidan bajo su impulso y preguntarse *para qué* estamos buscando el desarrollo de una IA que pueda llegar a emularnos. En una primera capa, encontrará la respuesta más evidente de la innegable mejora en el bienestar humano que produce la innovación tecnológica, a pesar de los daños colaterales que sin duda provoca. En una segunda capa, encontrará respuestas en torno a los intereses económicos en aprovechar el potencial de estas tecnologías, para fines dudosamente humanistas como los de maximizar nuestro consumo, optimizar la captación de nuestra atención e incluso alterar los procesos democráticos mediante la propaganda más eficiente de la historia. En capas más profundas, la prudencia humanista nos invita a reflexionar sobre la insatisfacción humana con su propia naturaleza, la soledad y el sinsentido profundos que experimenta el hombre en un mundo atomizado y secularizado y su interés en tratar de paliarlos con lenitivos virtuales y nuevas esperanzas transhumanistas de corte cuasi-religioso.

Junto a la prudencia, por otro lado, tendríamos a Pirrón y a Sexto Empírico con su ἐποχή, su *epojé* o suspensión del juicio, asociada a esa prudencia que invita a callar cuando no tenemos nada que decir, al estilo de Wittgenstein, a sobrevivir en la perplejidad y la incertidumbre ante nuevos fenómenos que aún tenemos que seguir investigando sin caer en fatalismos ni ingenuidades.

Cabe esperar en la cognición de la IA, aunque no parezca que la consciencia en sentido amplio esté cerca, ni que vaya a ser universal, sino más bien dependiente del modelo que se defina (Herzog & Herzog, 2024). En cualquier caso, el terreno pantanoso y altamente especulativo de la ciencia contemporánea no debe, no obstante, invitarnos a regresar cómodos a nuestra zona de confort para reafirmar nuestros prejuicios o nuestras viejas categorías. Es preferible asumir con entereza y madurez que a nuestro tiempo le ha tocado explorar nuevas respuestas humanistas ante el reto de la IA.

5. **CONCLUSIÓN**

En un contexto de inflación de expectativas sobre la IA, lo normal sería haber dedicado un artículo a bajar los humos a las grandes y triunfalistas proclamas de muchos sectores de la investigación. Especialmente a aquellos responsables de su marketing atentos a captar fondos, incluso aunque sea recabando la atención a costa del alarmismo social y de promesas inminentes que hoy son sólo ciencia ficción, como las de la singularidad de la superinteligencia o las del transhumanismo que promete llevarnos del carbono al silicio —o, acaso, al grafeno de propiedades cuánticas—. La IA, efectivamente, está muy lejos de ser lo que algunos de sus principales voceros proclaman ya, incluyendo la superación de la inteligencia humana. Sin embargo, en ciertos círculos de larga tradición humanista, esta reacción escéptica con los desarrollos tecnológicos más envalentonados viene de serie. Y parece aplaudir y congratularse cuando nuevos vientos anuncian posibles recensiones o nuevos inviernos en el desarrollo de la IA. Conviene, al contrario, procurarles una vacuna contra otro exceso: el de quien cree que es posible resolver con argumentos débiles y superados algunos de los desafíos y las reflexiones que los nuevos desarrollos tecnológicos están (re)abriendo. Eso es lo que se ha propuesto este artículo.

Cualquier propuesta humanista madura debe mantener su fidelidad y compromiso con la verdad, asumiendo con prudencia y apertura los conocimientos científicos más asentados y sus novedades sin olvidar su carácter provisional y acaso instrumental. Esto supone que, a pesar de la encendida defensa hecha aquí por reconocer el valor y por mantenerse al día de las novedades científicas más especializadas para pronunciarnos sobre cualquier asunto, como los avances de la IA hacia la inteligencia humana, esta propuesta no aboga por un cientifismo de fe ciega. La finitud epistémica humana es un hecho que palpa la propia ciencia pues, a su nivel más fundamental, Heisen-

berg y Gödel mediante, su propios principios y teoremas revelan un límite ineludible a la observación empírica y reconocen su incapacidad para autofundamentarse. Qué no decir del resto de ciencias elevadas sobre estos pies de barro más fundamentales, o de las limitaciones prácticas que el desarrollo tecnológico pueda enfrentar. Pero, para una pequeña luz que tenemos, procuremos no apagarla.

Además, este alegato por una aproximación humanista no reaccionaria no es ingenuo. Toda nueva tecnología lleva aparejado su accidente y su neutralidad moral es un mito ya desacreditado. Por ello, tendremos que saber protegernos ante los riesgos evidentes y ocultos en la IA, sin tener que llegar a los extremos de la singularidad ni del alarmismo excesivo. El carácter opaco de las conexiones internas a las redes neuronales profundas oculta sesgos que es preciso evidenciar, promoviendo una IA explicable (XAI) ante el reto de no erosionar los mejores rendimientos de los modelos menos transparentes. Por otro lado, el impacto de la IA en el tejido productivo podría poner en peligro importantes capas de la fuerza laboral a la que deberemos proteger con algún tipo de mecanismo similar a la renta universal si no se produce una creación equiparable de nuevos puestos de trabajo accesibles a la población y su nivel formativo. Además, la IA como herramienta potente para la generación de contenidos podría hacer, como en su día hizo la imprenta, que tiemblen los cimientos de la legitimidad que nos hemos dado en los últimos siglos y que las elecciones democráticas sufran importantes distorsiones, haciendo que unos hombres puedan imponer su voluntad sobre otros. La concentración de este poder tecnológico en pocas manos podría asimismo poner en peligro el equilibrio político y disparar las desigualdades sociales. Se otean riesgos sobre nuestra privacidad, nuestra integridad psicológica y moral, nuestra convivencia vulnerable a la desinformación. La IA podría servir para el desarrollo de armas letales autónomas o para asomarnos al abismo más estremecedor de nuestra propia reconfiguración genética.

Sin embargo, creo con esperanza que, mitigando sus impactos negativos, y manteniendo una postura humanista constructiva, cortés, prudente y abierta para tender puentes interdisciplinares, el desarrollo de la IA podrá seguir ayudándonos a descubrir aún más quiénes somos y a profundizar en nuestro desarrollo y crecimiento. Un humanismo no reaccionario ante la IA es posible.

Referencias

- Adams, S.; Arel, I., Bach, J.; Coop, R.; Furlan, R.; Goertzel, B.; Hall, J. S.; Samsonovich, A.; Scheutz, M.; Schlesinger, M.; Shapiro, S. C., y Sowa, J. (2012), Mapping the Landscape of Human-Level Artificial General Intelligence, en *AI Magazine* 33 (1), 25-42. https://doi.org/10.1609/aimag.v33i1.2322
- AlRegib, G., y Prabhushankar, M. (2022), Explanatory Paradigms in Neural Networks: Towards relevant and contextual explanations, en *IEEE Signal Processing Magazine* 39, 59-72. https://doi.org/10.1109/MSP.2022.3163871
- Appenzeller, T. (2018), Europe's first artists were Neandertals, en *Science, 359* (6378), 852-853. https://doi.org/10.1126/science.359.6378.852
- Aubert, M.; Brumm, A., y Huntley, J. (2018), Early dates for 'Neanderthal cave art' may be wrong, en *Journal of human evolution* 125, 215-217. https://doi.org/10.1016/j.jhevol.2018.08.004
- Baars, B. J. (2005), Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience, en *Progress in brain research* 150, 45-53.
- Beavan, A.; Domingo-Sananes, M. R., y McInerney, J. O. (2024), Contingency, repeatability, and predictability in the evolution of a prokaryotic pangenome, en *Proceedings of the National Academy of Sciences* 121 (1), e2304934120.
- Bekoff, M. (2000), Animal Emotions: Exploring Passionate Natures: Current interdisciplinary research provides compelling evidence that many animals experience such emotions as joy, fear, love, despair, and grief—we are not alone, en *BioScience* 50 (10), 861-870.
- Bengio, Y.; Lee, D.-H.; Bornschein, J.; Mesnard, T., y Lin, Z. (2016), Towards Biologically Plausible Deep Learning, en *ArXiv*: 1502.04156v3.
- Block, N. (1981), Psychologism and behaviorism, en *The Philosophical Review* 90 (1), 5-43.
- Bonhoeffer, D. (2001), *Resistencia y sumisión: cartas desde el cautiverio*, Salamanca: Sígueme.
- Brakes, P.; Dall, S. R. X.; Aplin, L. M.; Bearhop, S.; Carroll, E. L.; Ciucci, P.; Fishlock, V.; Ford, J. K. B.; Garland, E. C.; Keith, S. A.; McGregor, P. K.; Mesnick, S. L.; Noad, M. J.; Notarbartolo di Sciara, G.; Robbins, M. M.; Simmonds, M. P.; Spina, F.; Thornton, A.; Wade…, y Rutz, C. (2019), Animal cultures matter for conservation, en *Science* 363 (6431), 1032-1034.
- Bridges, A. D.; Royka, A.; Wilson, T.; Lockwood, C.; Richter, J.; Juusola, M., y Chittka, L. (2024), Bumblebees socially learn behaviour too complex to innovate alone, en *Nature*, 1-7.
- Brooks, T.; DePue, W.; Guo, Y.; Holmes, C.; Jing, L.; Luhman, E.; Luhman, T.; Ng, C.; Peebles, B.; Ramesh, A.; Schnurr, D.; Taylor, J., y Wang, R. (2024), Video generation models as world simulators. https://openai.com/research/video-generation-models-as-world-simulators

- Brown, R., Lau, H., y LeDoux, J. E. (2019), Understanding the higher-order approach to consciousness, en *Trends in cognitive sciences* 23 (9), 754-768.
- Burkhardt, A. (2022), El otro Wittgenstein o la "embestida contra los límites del lenguaje", en *Claridades: revista de filosofía* 14 (2), 101-140.
- Butlin, P.; Long, R.; Elmoznino, E.; Bengio, Y.; Birch, J.; Constant, A., *et al.* (2023), Consciousness in Artificial Intelligence: Insights from the Science of Consciousness, en *ArXiv* [cs.AI]. http://arxiv.org/abs/2308.08708
- Campolo A., y Crawford, K. (2020), Enchanted Determinism: Power Without Responsibility in Artificial Intelligence, en *Engaging Science, Technology, and Society* 6, 1-19.
- Chalmers, D. (1995), Facing up to the problem of consciousness, en *Journal of Consciousness Studies* 2 (3), 200-219. doi:10.1093/acprof:oso/9780195311105.003.0001
- Chalmers, D. (2007), The hard problem of consciousness, en *The Blackwell companion to consciousness*, 225-235.
- Chalmers, D. (2013), *La mente consciente*, Gedisa, Barcelona (ed. orig. 1996).
- Chalmers, D. (2018), The meta-problem of consciousness, en *Journal of Consciousness Studies* 25 (9-10), 6-61.
- Cuzzolin, F.; Morelli, A.; Cirstea, B., y Sahakian, B. (2020), Knowing me, knowing you: Theory of mind in AI, en *Psychological Medicine* 50, 1057-1061. https://doi.org/10.1017/S0033291720000835
- Damasio, A. (2012), *Y el cerebro creó al hombre*, Barcelona: Destino (ed. orig. 2010).
- Davis, E., y Marcus, G. (2015), Commonsense reasoning and commonsense knowledge in artificial intelligence, en *Communications of the ACM 58* (9), 92-103.
- Dehaene, S. (2015), *La conciencia en el cerebro*, Madrid: Siglo XXI (ed. orig. 2014).
- Dennett, D. C. (2013), *Intuition pumps and other tools for thinking*, WW Norton & Company.
- Dennett, D. C. (2018), Facing up to the hard question of consciousness, en *Philosophical Transactions of the Royal Society B: Biological Sciences,* 373 (1755), 20170342.
- Dobzhansky, T. (2013), Nothing in biology makes sense except in the light of evolution, en *The american biology teacher* 75 (2), 87-91.
- Dorobantu, M. (2022), Strong Artificial Intelligence and Theological Anthropology: One Problem, Two Solutions, en *Humanism and its Discontents: The Rise of Transhumanism and Posthumanism*, Cham: Springer International Publishing, 19-33.
- Dreyfus, H. (1972), *What Computers Can't Do*, New York: MIT Press.
- Fjelland, R. (2020), Why general artificial intelligence will not be realized. *Humanities and Social Sciences Communications, 7* (1), 1-9.

- Fraijó, M. (1998), *A vueltas con la religión*, Estella.
- Frankl, V. E. (2011), *El hombre en busca de sentido*, Madrid; Herder editorial (e. orig. 1946).
- Franks, N. R., y Richardson, T. (2006), Teaching in tandem-running ants, en *Nature, 439* (7073), 153-153.
- Freud, S. (1968), *Una dificultad del psicoanálisis*, en *Obras completas*, Madrid: Editorial Biblioteca Nueva.
- Gabrić, P. (2021), Overlooked evidence for semantic compositionality and signal reduction in wild chimpanzees (Pan troglodytes), en *Animal Cognition*, 1-13.
- Gaos, J. (1982), *Confesiones Profesionales*. *Aforística*, en *Obras Completas*, tomo XVII, México: UNAM.
- Greco, C. M., y Tagarelli, A. (2023), Bringing order into the realm of Transformer-based language models for artificial intelligence and law, en *Artificial Intelligence and Law*, 1-148.
- Griffiths, T. (2020), Understanding Human Intelligence through Human Limitations. *Trends in Cognitive Sciences, 24*, 873-883. https://doi.org/10.1016/j.tics.2020.09.00
- Guerrero, L. E.; Castillo, L. F.; Arango-López, J., y Moreira, F. (2023), A systematic review of integrated information theory: a perspective from artificial intelligence and the cognitive sciences, en *Neural Computing and Applications*, 1-33.
- Guha, N.; Nyarko, J.; Ho, D. E.; Ré, C.; Chilton, A.; Narayana, A.; Chohlas-Wood, A.; Peters, A.; Waldon, B.; Rockmore, D. N.; Zambrano, D.; Talisman, D.; Hoque, E.; Surani, F.; Fagan, F.; Sarfaty, G.; Dickinson, G. M.; Porat, H.: Hegland, J…, y Li, Z. (2023), Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models, en *ArXiv* preprint *ArXiv:2308.11462*.
- Haikonen, P. O. (2020), On artificial intelligence and consciousness, en *Journal of Artificial Intelligence and Consciousness* 7 (01), 73-82.
- Hameroff, S., y Penrose, R. (1996), Conscious Events as Orchestrated Space-Time Selections, en *Journal of Consciousness Studies* 3 (1), 36–53. http://www.ingentaconnect.com/content/imp/jcs/1996/00000003/00000001/679.
- Hameroff, S., y Penrose, R. (2014), Consciousness in the universe: A review of the 'Orch OR'theory, en *Physics of life reviews* 11 (1), 39-78.
- Harnad, S. (1990), The Symbol Grounding Problem, en *Physica D: Nonlinear Phenomena* 42 (1-3), 335-346.
- Hassabis, D.; Kumaran, D.; Summerfield, C., y Botvinick, M. (2017), Neuroscience-Inspired Artificial Intelligence, en *Neuron* 95, 245-258.
- Hefner, P. (2019), Biocultural evolution and the created co-Creator, en *Science and Theology*, editado por Ted Peters, Routledge,174-188.
- Herzog, D. J., y Herzog, N. (2024), *What is it like to be an AI bat?*, *Qeios*.
- Hofstadter, D. R. (1999), *Gödel, Escher, Bach: An eternal golden braid*, Basic books.

- Hume, D. (2000), *A Treatise of Human Nature*, David Fate Norton y Mary J. Norton (Ed.), Oxford University Press, (ed. orig. 1739).
- Jackson, F. (1998), Epiphenomenal qualia, en *Consciousness and emotion in cognitive science,* Routledge, 197-206.
- Janik, V. M. (2013), Cognitive skills in bottlenose dolphin communication, en *Trends in cognitive sciences* 17 (4), 157-159.
- Kaspar, C.; Ravoo, B.; Wiel, W.; Wegner, S., y Pernice, W. (2021), The rise of intelligent matter, en *Nature* 594, 345-355. https://doi.org/10.1038/s41586-021-03453-y
- Kim, S. Y.; Schmitt, B. H., y Thalmann, N. M. (2019), Eliza in the uncanny valley: Anthropomorphizing consumer robots increases their perceived warmth but decreases liking, en *Marketing letters* 30, 1-12.
- Kirk, R. (2005). *Zombies and consciousness*, Clarendon Press.
- Koch, C.; Massimini, M.; Boly, M., y Tononi, G. (2016), Neural correlates of consciousness: progress and problems, en *Nature Reviews Neuroscience* 17 (5), 307-321.
- Kurzweil, R. (2005), *The Singularity is Near: When Humans Transcend Biology*, Viking Penguin.
- Larson, E. J. (2021), *The Myth of Artificial Intelligence: Why computers can't think the way we do*, Harvard University Press.
- Lee, J. P.; Jang, H.; Jang, Y.; Song, H.; Lee, S.; Lee, P. S., y Kim, J. (2024), Encoding of multi-modal emotional information via personalized skin-integrated wireless facial interface, en *Nature Communications* 15 (1), 530.
- Lenski, R. E. (2023), Revisiting the design of the long-term evolution experiment with Escherichia coli, en *Journal of Molecular Evolution,* 1-13.
- Levine, J. (1983), Materialism and qualia: The explanatory gap, en *Pacific Philosophical Quarterly* 64, no. 4, 354-361. doi:10.1111/j.1468-0114.1983.tb00207.x.
- Lindell, A. (2013), Continuities in Emotion Lateralization in Human and Non-Human Primates, en *Frontiers in Human Neuroscience* 7. https://doi.org/10.3389/fnhum.2013.00464
- Llinás, R. (2003), *El cerebro y el mito del yo,* Cali: Norma, (ed. orig. 2002).
- Locke, J. (1975), *Essay Concerning Human Understanding*, Oxford: Oxford University Press (ed. orig. 1689).
- Lucas, J. (1996), Minds, machines and Gödel: A retrospect, en *Artificial intelligence: Critical concepts* 3, 359-376.
- Lumbreras, S. (2017), Strong artificial intelligence and imago hominis: The risks of a reductionist definition of human nature, en *Issues in Science and Theology: Are We Special? Human Uniqueness in Science and Theology*, 157-168.
- Lumbreras, S. (2022), Lessons from the quest for Artificial Consciousness: The emergence criterion, insight-oriented AI, and Imago Dei, en *Zygon* 57, 963-983. https://doi.org/10.1111/zygo.12827

- Lyre, H. (2020), The state space of artificial intelligence, en *Minds and Machines* 30 (3), 325-347.
- MacDonald, K.; Scherjon, F.; van Veen, E.; Vaesen, K., y Roebroeks, W. (2021), Middle Pleistocene fire use: The first signal of widespread cultural diffusion in human evolution, en *Proceedings of the National Academy of Sciences* 118 (31).
- Madrid, C. (2024), *Filosofía de la inteligencia artificial*, Oviedo: Pentalfa.
- Manzotti, R. (2019), Mind-object identity: A solution to the hard problem, en *Frontiers in Psychology* 10 (FEB), art. no. 63. https://doi.org/10.3389/fpsyg.2019.00063.
- Manzotti, R. (2021), The boundaries and location of consciousness as identity theories deem fit [I confini e la localizzazione della coscienza secondo le teorie dell'identità], en *Rivista Internazionale di Filosofia e Psicologia* 12 (3), 225-241. https://doi.org/10.4453/rifp.2021.0022.
- Mascaro *et al.* (2022), Application of insects to wounds of self and others by chimpanzees in the wild, en *Current Biology* 32 (3), R112-R113.
- McGinn, C. (1989), Can we solve the Mind–Body problem?, en *Mind* 98 (391), 349-366.
- Metzinger, T. (2013), Two principles for robot ethics, en *Robotik und gesetzgebung*, 247-286.
- Mingers, J. (2012), Abduction: The missing link between deduction and induction. A comment on Ormerod's "rational inference: Deductive, inductive and probabilistic thinking", en *Journal of the Operational Research Society* 63, 860-861. https://doi.org/10.1057/jors.2011.85.
- Mitchell, M. (2019), *Artificial intelligence: A guide for thinking humans*, Penguin UK.
- Mora, F. (2005), *El reloj de la sabiduría: tiempos y espacios en el cerebrohumano,* Madrid: Alianza.
- Mosterín, J. (1998), *Vivan los animales.* Debate S.A.
- Nagel, T. (1980), What is it like to be a bat?, en *The Language and Thought Series*, Harvard University Press, 159-168.
- Nesse, R., y Ellsworth, P. (2009), Evolution, emotions, and emotional disorders, en *The American psychologist* 64 (2), 129-139. https://doi.org/10.1037/a0013503
- Newen, A.; De Bruin, L., y Gallagher, S. (2018), The Oxford Handbook of 4E Cognition, Oxford, UK: Oxford University Press.
- Niikawa, T. (2020), A map of consciousness studies: questions and approaches, en *Frontiers in Psychology* 11, 530152.
- Osuna-Mascaró, A. J. (2022), Innovative composite tool use by Goffin's cockatoos (Cacatua goffiniana), en *Scientific Reports* 12 (1), 1-10.
- Oviedo, L. (2022), Artificial Intelligence and Theology: Looking for a Positive —but Not Uncritical— Reception, en *Zygon* 57 (4), 938-952.

- Pascal, B. (2015), *Pensamientos* (trad. Javier Zubiri), Madrid: Alianza Editorial, (ed. orig. 1669).
- Peebles, W., y Xie, S. (2023), Scalable diffusion models with transformers, en *Proceedings of the IEEE/CVF International Conference on Computer Vision* 4195-4205.
- Peirce, C. S. (1992), *The Essential Peirce, Volume 2: Selected Philosophical Writings (1893-1913)* (vol. 2), N. Houser *et al.* (eds.), Bloomington: Indiana University Press.
- Pfandler, A.; Rümmele, S., y Szeider, S. (2013), Backdoors to Abduction, en *ArXiv*, abs/1304.5961.
- Pinker, S. (2007), The mystery of consciousness, en *Time* 169 (5), 58-62.
- Plebe, A., y Perconti, P. (2020), Plurality: The End of Singularity?, en Korotayev, A., y LePoire, D. (Eds), *The 21st Century Singularity and Global Futures. World-Systems Evolution and Global Futures*. Springer, Cham. https://doi.org/10.1007/978-3-030-33730-8_8
- Pugh, G. E. (1977), *The Biological Origin of Human Values*, New York: Basic Books.
- Putnam, H. (1975), The nature of mental states, en *Philosophical Papers*, Cambridge: Cambridge University Press, 429-440.
- Roth, G., y Dicke, U. (2005), Evolution of the brain and intelligence, en *Trends in Cognitive Sciences* 9, 250-257. https://doi.org/10.1016/j.tics.2005.03.005
- Sánchez-Cañizares, J. (2014), The Mind-Brain Problem and the Measurement Paradox of Quantum Mechanics: Should We Disentangle Them?, en *Neuro-Quantology* 12 (1): 76-95. doi:10.14704/nq.2014.12.1.696
- Sánchez-Cañizares, J. (2016), Neurociencia y mecánica cuántica, en *Diccionario Interdisciplinar Austral*, editado por Claudia E. Vanney, Ignacio Silva y Juan F. Franck. http://dia.austral.edu.ar/Neurociencia_y_mecánica_cuántica
- Searle, J. R. (1980), Minds, brains, and programs, en *Behavioral and brain sciences* 3 (3), 417-424.
- Searle, J. R. (2006), Chinese room argument, en *Scholarpedia* 4, 3100. https://doi.org/10.1002/0470018860.S00159
- Sejnowski, T. J. (2018), *The Deep Learning Revolution,* MIT Press.
- Sequeiros, L. (2023), "Tender puentes" versus "apologética": dos estrategias en el encuentro ciencia-religión, en *Fronteras CTR*. https://blogs.comillas.edu/FronterasCTR/?p=8087
- Seth, A. K., y Bayne, T. (2022), Theories of consciousness, en *Nature Reviews Neuroscience* 23 (7), 439-452.
- Simondon, G. (2007), *El modo de existencia de los objetos técnicos*, Prometeo Libros Editorial.
- Smith, D., y Schillaci, G. (2021), Why Build a Robot With Artificial Consciousness? How to Begin? A Cross-Disciplinary Dialogue on the Design and Imple-

mentation of a Synthetic Model of Consciousness, en *Frontiers in Psychology* 12, 1107. https://doi.org/10.3389/fpsyg.2021.530560
– Surden, H. (2019), Artificial Intelligence and Law: An Overview. *Georgia State University Law Review*, *35*, 19-22. https://ssrn.com/abstract=3411869.
– Taddeo, M., y Floridi, L. (2005), Solving the symbol grounding problem: A critical review of fifteen years of research, en *J Experimental Theoretical Artifi Intell* 17 (4), 419-445.
– Tait, I.; Bensemann, J., y Nguyen, T. (2023), Building the Blocks of Being: The Attributes and Qualities Required for Consciousness, en *Philosophies* 8 (4), 52.
– Tatay, J. (2023). François Euvé (2022), La science l'épreuve de Dieu?, en *Razón y Fe* 287 (1462), 315-317. Recuperado a partir de https://revistas.comillas.edu/index.php/razonyfe/article/view/20671
– Tevet, O.; Gross, R. D.; Hodassman, S.; Rogachevsky, T.; Tzach, Y.; Meir, Y., y Kanter, I. (2024), Efficient shallow learning mechanism as an alternative to deep learning, en *Physica A: Statistical Mechanics and its Applications*, vol. 635, 129513.
– Turing, A. M. (1950), Computing Machinery and Intelligence, en *Mind.* 49 (236): 433-460. doi:10.1093/mind/LIX.236.433.
– Turing, A. M. (1951), Can Digital Computers Think?, en B. J. Copeland (ed.), *The Essential Turing* (Oxford, 2004, online edn, Oxford Academic, 12 Nov. 2020), https://doi.org/10.1093/oso/9780198250791.003.0019
– Udías, A. (2014), *Los jesuitas y la ciencia. Una tradición en la Iglesia,* Bilbao: Ediciones Mensajero.
– Ullman, S. (2019), Using neuroscience to develop artificial intelligence, en *Science* 363 (6428), 692-693.
– Unamuno, M. (1913), Mecanópolis, *El imparcial*.
– Waterhouse, L. (2023), Why multiple intelligences theory is a neuromyth, en *Frontiers in Psychology* 14.
– Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J., y Fedus, W. (2022), Emergent abilities of large language models, en *ArXiv* preprint *ArXiv:2206.07682*.
– Williams, J.; Fiore, S., y Jentsch, F. (2022), Supporting Artificial Social Intelligence With Theory of Mind, en *Frontiers in Artificial Intelligence* 5. https://doi.org/10.3389/frai.2022.750763
– Zednik, C. (2019), Solving the black box problem: A normative framework for explainable artificial intelligence, en *Philos Technol.* https://doi.org/10.1007/s13347-019-00382-7
- Zhuang, Y.; Cai, M.; Li, X.; Luo, X.; Yang, Q., y Wu, F. (2020), The Next Breakthroughs of Artificial Intelligence: The Interdisciplinary Nature of AI, en *Engineering* 6 (3), 245-247. https://doi.org/10.1016/j.eng.2020.01.009